TESIS DE MAESTRÍA EN CIENCIAS FÍSICAS

ANÁLISIS ESTADÍSTICO DE LA VARIACIÓN POBLACIONAL DE LAS CARACTERÍSTICAS ESTRUCTURALES DEL ENCÉFALO HUMANO

Juan Bautista Arenaza Manzo

Inés Samengo
Directora

Jurado
Damián Hernández
Luis Moyano
Diego Fernández Slezak

22 de Febrero de 2021

Departamento de Física Médica – Centro Atómico Bariloche

Instituto Balseiro
Universidad Nacional de Cuyo
Comisión Nacional de Energía Atómica
Argentina

INVENTARIO: 24139
12.07.2021
Biblioteca Leo Falicov

Índice de contenidos

Resumen

La relación entre la experiencia de vida de un dado individuo y sus características neuroanátomicas es un tema muy controversial. Frecuentemente se discute la importancia del tamaño de determinada estructura cerebral en las capacidades o el comportamiento de las personas. La postura que toma una sociedad en estos temas puede tener grandes consecuencias, como por ejemplo la estigmatización de algunos de sus individuos. En este contexto se vuelve de vital importancia que las afirmaciones sobre las relaciones entre aspectos socioambientales y neuroanatómicos estén sustentadas por suficiente evidencia. Por este motivo, hicimos un relevamiento de tanto características sociales como neuroanatómicas de una muestra de 193 individuos adultos.

El estudio conlleva el diseño de procedimientos bayesianos para reducir la dimensión, en el que detectamos conjuntos cuasi-independientes de medidas cerebrales que co-varían fuertemente, y en consecuencia, no necesitan describirse de manera independiente. Cada uno de estos conjuntos contiene todas las medidas de unas pocas estructuras, lo cual nos permitió describirlos a través de variables latentes fácilmente interpretables.

Posteriormente estimamos las relaciones entre variables latentes y socioambientales, permitiendo que las variables latentes interactúen entre sí. El algoritmo de inferencia es conservador, en el sentido que estima únicamente las relaciones significativamente distintas de cero. Encontramos que los efectos socioambientales más robusto involucran al sexo, la altura y el número de embarazos.

Capítulo 1

Motivación

Los medios frecuentemente presentan afirmaciones polémicas acerca de la relevancia del tamaño de determinadas áreas cerebrales en las capacidades intelectuales o morales de las personas. Ejemplos clásicos son: las mujeres tienen el cerebro más chico que los hombres [1], o los criminales tienen volúmenes reducidos en áreas relacionadas con funciones ejecutivas, control emocional y toma de decisiones, y simultáneamente tienen volúmenes aumentados en las regiones asociadas a recompensa [2]. Estas afirmaciones tienen consecuencias importantes en los medios de comunicación, y en la estigmatización de individuos, y si son infundadas, rayan en la frenología. Esto no es motivo, sin embargo, para abstenerse de hacer toda afirmación. Lo importante es que sólo se afirme la existencia de correlaciones entre factores socioambientales y factores neuroanatómicos solo en tanto y en cuanto existe evidencia suficiente para hacer la conexión. En esta tesis implementamos algoritmos bayesianos para inferir la estructura de correlaciones entre variables neuroanatómicas entre sí, y entre variables anatómicas y variables socioambientales. Ponemos el énfasis en (a) utilizar criterios objetivos y conservadores para determinar la relevancia de una correlación, (b) obtener resultados que sean interpretables en términos anatómicos, y (c) poner el foco en la existencia o no de correlaciones - sin inferir de ellas relaciones causales.

Capítulo 2

Introducción

2.1. Métodos

2.1.1. Imágenes por resonancia magnética nuclear

La resonancia magnética nuclear es una técnica utilizada para obtener imágenes del interior del cuerpo y los procesos que ocurren en él sin usar métodos invasivos. Las imágenes se generan mediante campos magnéticos oscilantes y ondas de radio que interactúan excitando protones de hidrógeno. La frecuencia de oscilación entre estados de equilibrio y excitados permite diferenciar tipos de tejido y generar imágenes en base a esta información. En este estudio medimos características estructurales del cerebro humano a partir del procesamiento de imágenes por resonancia magnética nuclear.

2.1.2. Resonador y secuencias utilizadas

Se usaron dos resonadores para la adquisición de imágenes en este estudio. En Florencio Varela (Buenos Aires) se utlizó el resonador Philips Achieva de 3 T ubicado en el hospital El Cruce. En San Carlos de Bariloche escaneamos a los participantes con el equipo SIGNA PET/MR de 3 T de GE Healthcare que se encuentra en el Centro de Radioterapia y Medicina Nuclear de Bariloche. A partir de cada escaneo obtuvimos tres imágenes: una anatómica T1 volumétrica, una de difusión y una funcional de 250 volúmenes. Una resonancia completa tenía una duración de 30 minutos, durante los cuales el sujeto se encontraba en estado de reposo y despierto.

En esta tesis de maestría nos concentramos en estudiar las características estructurales del cerebro humano a partir de la información que obtuvimos de las imágenes T1, dejando el análisis de las imágenes de difusión y funcionales para trabajo futuro. En la Fig. 2.1 mostramos cortes en los planos axial, sagital y coronal de una imagen anatómica obtenida mediante una secuencia T1.

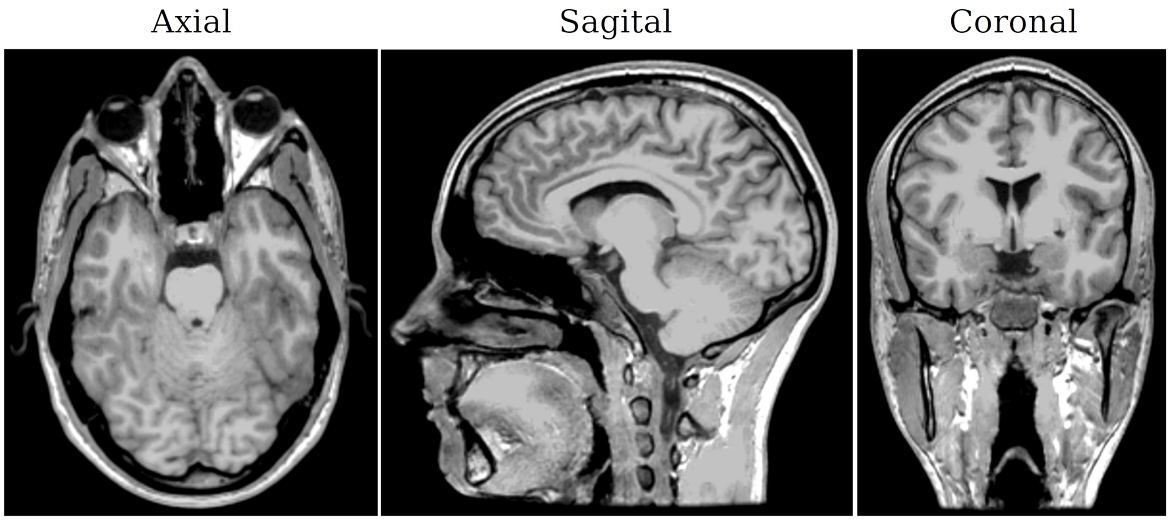

Figura 2.1: Ejemplos de imágenes obtenidas en los escaneos. Mostramos cortes en los planos axial, sagital y coronal.

2.1.3. Segmentación

La identificación de estructuras cerebrales a partir de imágenes T1 se realizó con el programa libre FreeSurfer. Brevemente, el proceso consiste en: corregir errores por movimiento [3], eliminar tejidos no cerebrales [4], normalizar la intensidad [5], hacer un registro a un atlas segmentado [6–8], detectar estructuras de sustancia blanca y gris en la imagen, y tomar medidas de volumen, área o espesor para cada una de ellas [9]. Cada parcelación se hizo a partir del atlas Desikan-Killiany [9] (ver Fig. 2.2), el cual brinda medidas de volumen, área y espesor de ambos hemisferios de 31 estructuras corticales. Además, el programa ofrece medidas de volumen de materia blanca y de estructuras subcorticales. En total, cada segmentación nos proporcionaba $d_a = 294$ medidas anatómicas. A modo de ejemplo, en la Fig. 2.3 mostramos una segmentación hecha por el FreeSurfer.

2.1.4. Población muestreada

La población objeto de estudio está conformada por individuos adultos de ambos sexos, sin antecedentes de enfermedades neurológicas o psiquiátricas, residentes en San Carlos de Bariloche, Florencio Varela (Buenos Aires), y sus respectivas zonas de influencia. Al mismo tiempo, excluimos aquellos sujetos que hayan presentado hallazgos incidentales, definidos como rasgos atípicos detectados en un relevamiento de las imágenes por dos médicos: el neurólogo Dr. Sergio Lindenbaum en el caso de la muestra de Bariloche y el especialista en imágenes Dr. Juan Pablo Princich para la muestra de Florencio Varela. Si bien hasta el momento de la participación en el experimento tales anomalías no tuvieron una manifestación clínica que motivara una consulta médica, su discrepancia con los parámetros considerados normales justificó la exclusión de la po-

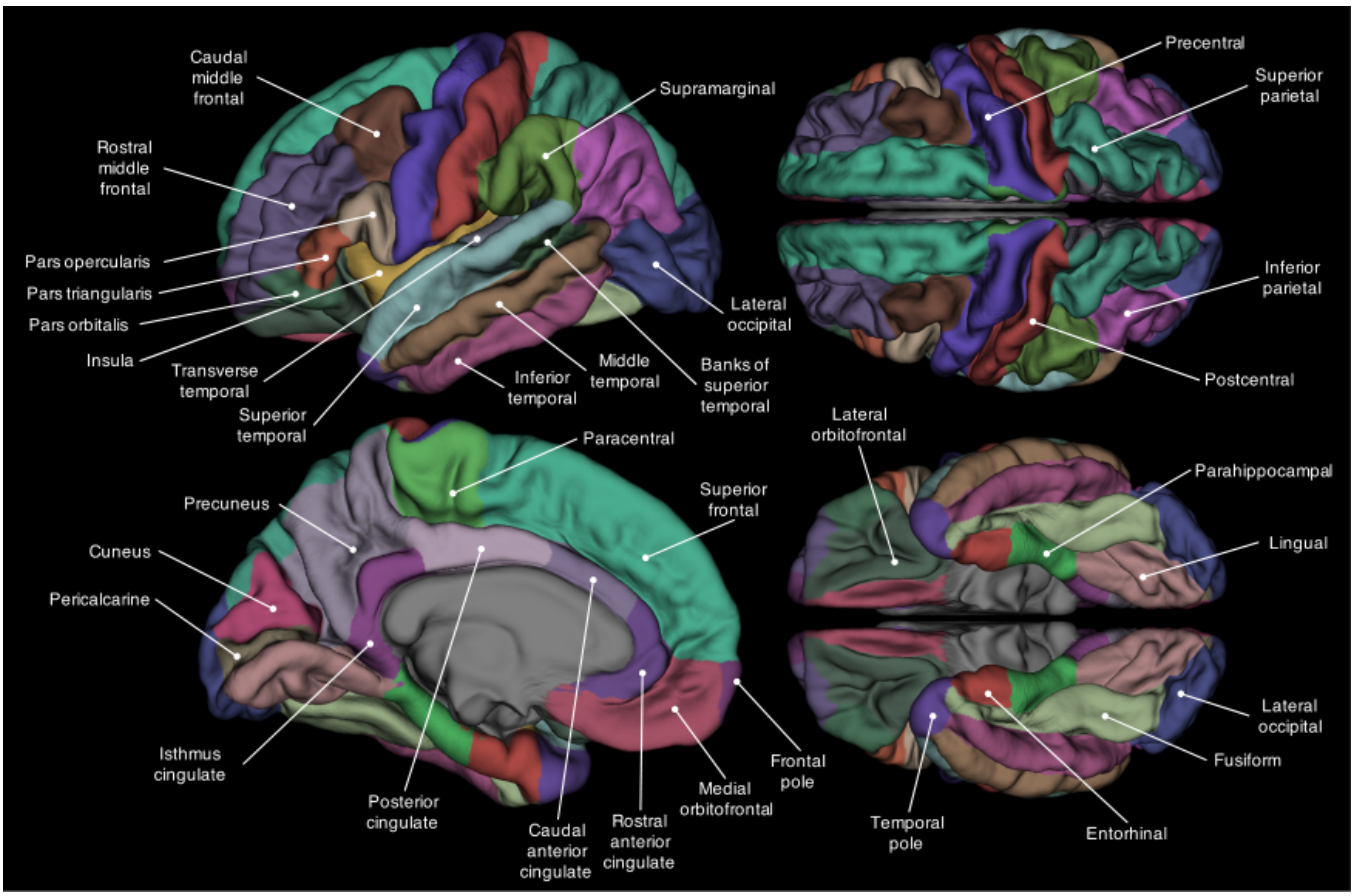

Figura 2.2: Estructuras corticales identificadas en el atlas Desikan-Killiany.

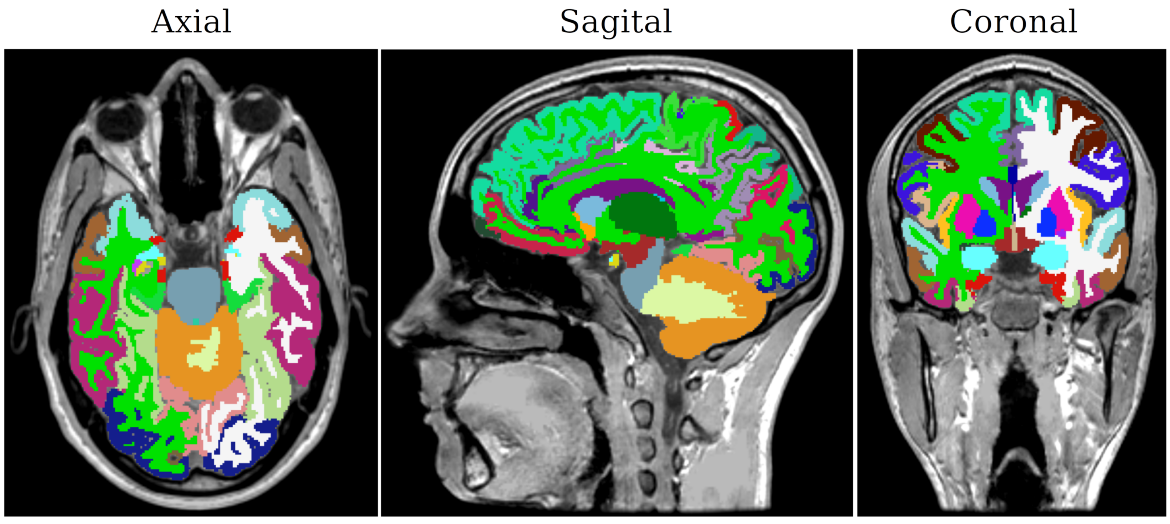

Figura 2.3: Ejemplo de una segmentación realizada por el programa FreeSurfer a partir del atlas Desikan-Killiany.

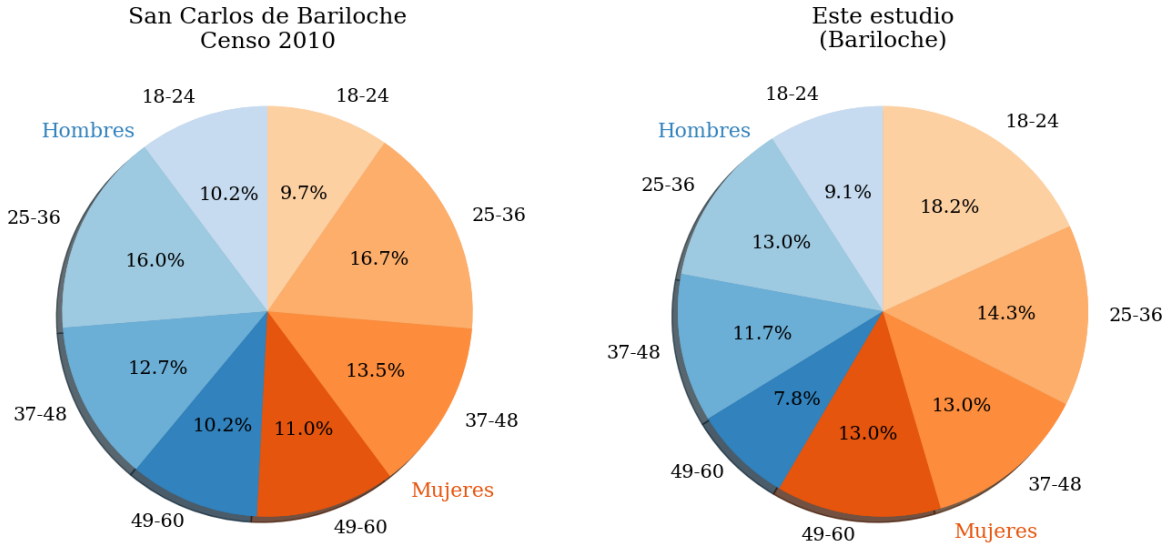

Figura 2.4: Estructura poblacional de San Carlos de Bariloche (izquierda) y de la población barilochense muestreada para este estudio (derecha). Se muestra el porcentaje de personas por rango etario y sexo. Los datos del Censo 2010 nos fueron brindados por la Delegación Regional INDEC Patagonia.

blación control. Llevamos a cabo el reclutamiento de personas considerando el carácter estrictamente voluntario de la participación. Cada participante completa una encuesta socioambiental (se adjunta en el apéndice A) y firma un consentimiento informado. Este proyecto fue aprobado por el comité de ética del Hospital Néstor Kirchner y cuenta con financiación de organismos nacionales (CONICET, FonCyT). Juntando los datos adquiridos en los resonadores de San Carlos de Bariloche y Florencio Varela, nuestra muestra es de $n = 193$ participantes.

Reclutamiento en San Carlos de Bariloche

Tanto el reclutamiento de participantes como la adquisición de imágenes en San Carlos de Bariloche forman parte de este trabajo de maestría. Con el objetivo de que la población muestreada sea representativa de la región de interés, solicitamos datos de la estructura poblacional de San Carlos de Bariloche a la Delegación Regional INDEC Patagonia. La Fig. 2.4 muestra los datos poblacionales del Censo 2010 y los correspondientes a los 77 voluntarios (32 hombres y 45 mujeres) de San Carlos de Bariloche que participaron en nuestro estudio. Observamos que, salvo quizás por la proporción de mujeres entre 18 y 24 años, la composición etaria y por sexo de nuestra muestra es similar a la de la población objeto de estudio.

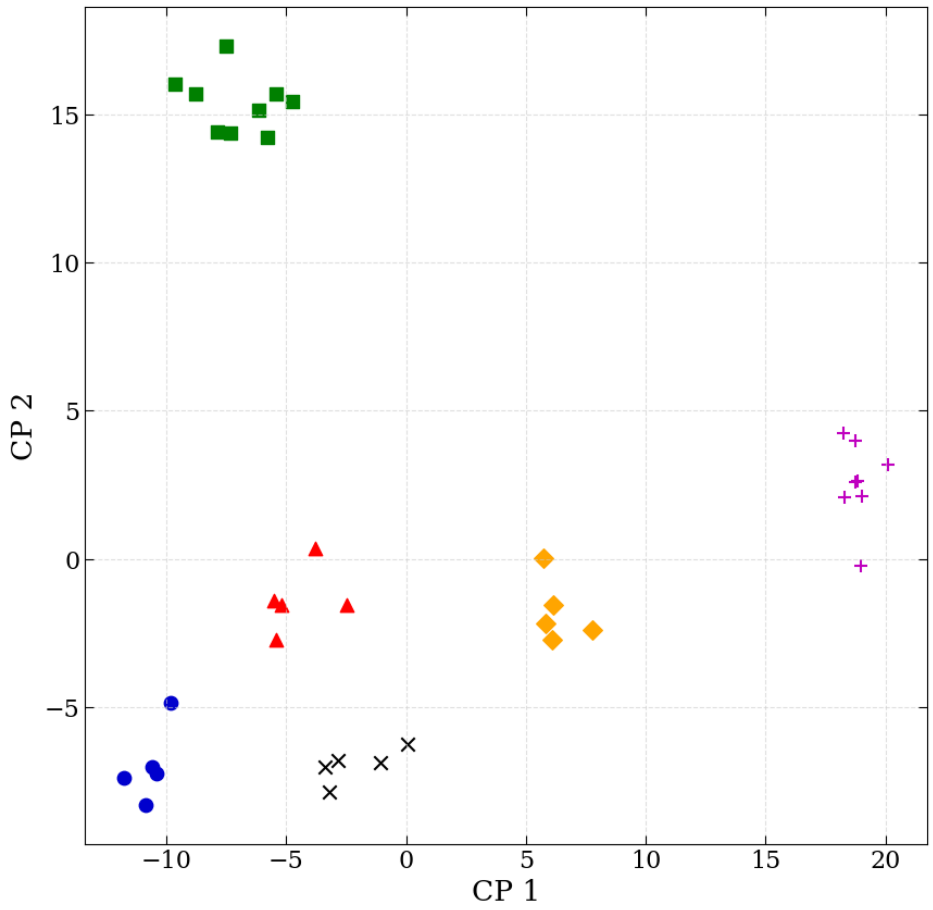

Figura 2.5: Primeras dos componentes principales (CP) de los datos correspondientes a los sujetos escaneados múltiples veces. Puntos de distinta forma y color implican sujetos distintos. Pre-procesamos los datos corriendo el origen a la media poblacional y normalizando cada coordenada por su desviación estándar.

2.1.5. Adquisiciones repetidas

Con el objetivo de caracterizar la variabilidad intrasujeto y determinar su relación con la intersujeto, escaneamos múltiples veces a seis participantes. Cuatro de ellos se hicieron cinco resonancias, uno se hizo ocho, y el restante nueve. Para visualizar las adquisiciones repetidas rápidamente, en la Fig. 2.5 mostramos sus primeras dos componentes principales, donde pre-procesamos los datos corriendo el origen a la media poblacional y normalizando cada coordenada por su desviación estándar. A pesar de que existe cierta variabilidad intrasujeto en el subespacio de las primeras dos componentes principales, parece ser lo suficientemente pequeña como para no confundirla con la intersujeto.

2.2. Análisis preliminar

A continuación argumentamos la importancia de buscar un subespacio de tamaño menor al original donde representar los datos, y describimos los resultados obtenidos

mediante un análisis de componentes principales (PCA) sobre los datos de la población de San Carlos de Bariloche.

2.2.1. Importancia de reducir la dimensión

La importancia de la selección del subespacio donde representar los datos radica en que influye en la significancia de la discriminación de grupos de personas por factores socioambientales. En las Figs. 2.6 y 2.7 mostramos los resultados de dos simulaciones que apoyan este argumento. En ambas simulaciones evaluamos la significancia de la discriminación entre dos conjuntos X e Y de vectores d-dimensionales mediante el factor de Bayes, definido como

$$\mathrm{FB} = \frac{P(X,Y|H_1)}{P(X,Y|H_0)},$$

donde H_1 y H_0 son las hipótesis "X e Y salen de distribuciones gaussianas distintas" y "X e Y salen de una misma distribución gaussiana" respectivamente. Es decir que

$$\begin{aligned} P(X,Y|H_1) &= P(X|H_1)\,P(Y|H_1) \\ &= \left\{ \int \mathrm{d}\boldsymbol{\mu}_X\,\mathrm{d}C_X\; P(\boldsymbol{\mu}_X, C_X|H_1) \prod_{\boldsymbol{x}\in X} \frac{e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_X)^T C_X^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_X)}}{\left[(2\pi)^d \det(C_X)\right]^{1/2}} \right\} \\ &\quad \times \left\{ \int \mathrm{d}\boldsymbol{\mu}_Y\,\mathrm{d}C_Y\; P(\boldsymbol{\mu}_Y, C_Y|H_1) \prod_{\boldsymbol{y}\in Y} \frac{e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_Y)^T C_Y^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_Y)}}{\left[(2\pi)^d \det(C_Y)\right]^{1/2}} \right\} \end{aligned}$$

y

$$P(X,Y|H_0) = \int \mathrm{d}\boldsymbol{\mu}\,\mathrm{d}C\; P(\boldsymbol{\mu}, C|H_0) \prod_{\boldsymbol{z}\in X\cup Y} \frac{e^{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T C^{-1}(\boldsymbol{z}-\boldsymbol{\mu})}}{\left[(2\pi)^d \det(C)\right]^{1/2}},$$

donde usamos distribuciones a priori $P(\boldsymbol{\mu}_X, C_X|H_1)$, $P(\boldsymbol{\mu}_Y, C_Y|H_1)$ y $P(\boldsymbol{\mu}, C|H_0)$ conjugadas y poco informativas (anchas).

En la simulación correspondiente a la Fig. 2.6 partimos de dos conjuntos de datos representados por vectores d-dimensionales, cuya primera componente es generada con dos distribuciones de probabilidad distintas, pero las componentes subsiguientes provienen de la misma distribución. Al calcular el factor de Bayes utilizando la dimensión relevante y un número variable de dimensiones irrelevantes, la significancia se pierde rápidamente. Para la simulación generamos dos conjuntos de datos $X_0 = \{x_1, \ldots, x_{100}\}$ e $Y_0 = \{y_1, \ldots, y_{100}\}$ a partir de distribuciones normales unidimensionales: $x_i \sim \mathcal{N}(-2, 1)$ e $y_i \sim \mathcal{N}(2, 1)$, y computamos el factor de Bayes para comparar las hipótesis: "X_0 e Y_0 salen de distribuciones distintas" o "X_0 e Y_0 salen de una misma distribución". A continuación, agregamos una dimensión irrelevante al espacio de los datos generados,

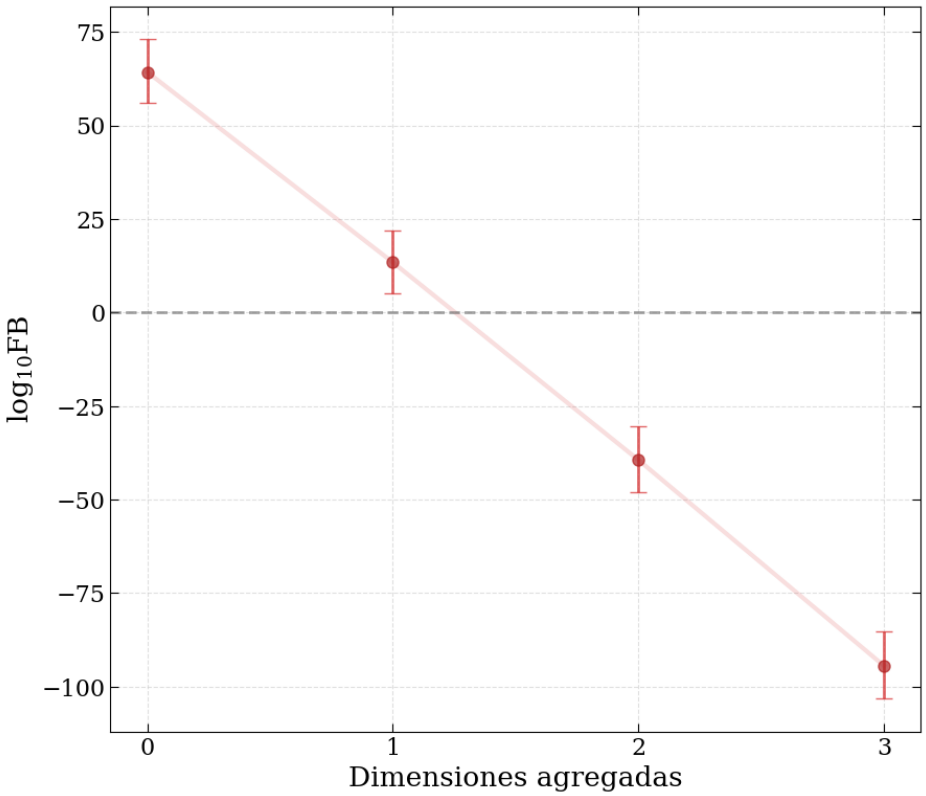

Figura 2.6: Logaritmo del factor de Bayes en función de las dimensiones irrelevantes agregadas para el caso en que los datos se muestrean de distribuciones distintas. Los puntos mostrados corresponden a la media de las distribuciones estimadas y las barras de incerteza contienen al 95 % de las muestras. Vemos que si el número de dimensiones irrelevantes es mayor al de dimensiones relevantes, el factor de Bayes deja de favorecer la hipótesis de que las distribuciones son distintas.

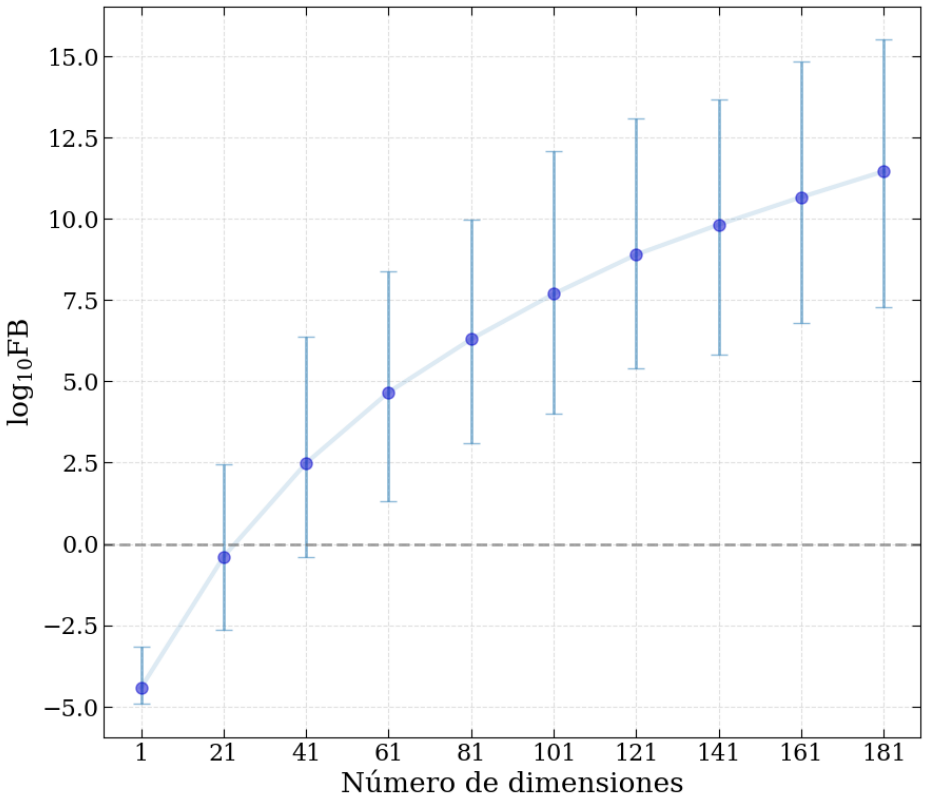

Figura 2.7: Logaritmo del factor de Bayes sobre la dirección que maximiza la discriminación de dos conjuntos de datos generados a partir de una misma distribución en función del número de dimensiones del espacio original. Los puntos mostrados corresponden a la media de las distribuciones estimadas y las barras de incerteza contienen al 95 % de las muestras. Cuando la dimensión es mayor que el número de datos, el factor de Bayes favorece la hipótesis de que los dos conjuntos provienen de distribuciones distintas.

es decir, a cada dato le añadimos una componente $z_i \sim \mathcal{N}(0, 1)$. Entonces, los conjuntos de datos quedan $X_1 = \{(x_1, z_1), \ldots, (x_{100}, z_{100})\}$ e $Y_1 = \{(y_1, z_{101}), \ldots, (y_{100}, z_{200}\}$. Continuamos agregando dimensiones irrelevantes y, en cada iteración, computamos el factor de Bayes para evaluar la significancia de la discriminación de los conjuntos X_d e Y_d. Repetimos este procedimiento 1000 veces para estimar la distribución de $\log_{10}$FB en función del número de dimensiones agregadas. En la Fig. 2.6 mostramos los resultados obtenidos, donde se observa que cada dimensión irrelevante agregada disminuye la significancia estimada por el factor de Bayes.

Al mismo tiempo, la Fig. 2.7 muestra que, si partimos de conjuntos de datos que salen de una misma distribución de probabilidad, basta con aumentar la dimensión del espacio para que aparezca una dirección sobre la cual los conjuntos parezcan significativamente distintos. Para cada número de dimensiones $d \in \{1, 21, 41, \ldots, 181\}$, generamos dos conjuntos de datos $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{20}\}$ e $Y = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{20}\}$ a partir de una misma distribución normal: $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \mathbb{1})$ e $\boldsymbol{y}_i \sim \mathcal{N}(\boldsymbol{0}, \mathbb{1})$, buscamos la dirección que maximiza la discriminación entre tales conjuntos y computamos el factor de Bayes. Al igual que en la simulación anterior, repetimos este proceso 1000 veces para estimar

la distribución de $\log_{10}$FB en función del número de dimensiones del espacio generado. En la Fig. 2.7 mostramos los resultados y vemos que, si no descartamos dimensiones cuidadosamente, podemos generar una falsa significancia en la distinción de dos conjuntos de datos.

2.2.2. Análisis de componentes principales

Uno de los algoritmos de reducción de dimensión más utilizados es el análisis de componentes principales o PCA (*principal component analysis*). Dado un conjunto de datos observados, PCA permite computar las variables que generan el subespacio en el que la varianza retenida es máxima. Estas nuevas variables se conocen como componentes principales (CP) y corresponden a un cambio de base, es decir, cada CP es combinación lineal de todas las variables originales.

Dado que las medidas anatómicas brindadas por el programa FreeSurfer no son dimensionalmente homogéneas y tienen distinto orden de magnitud, antes de aplicar PCA pre-procesmos los datos de la población de San Carlos de Bariloche corriendo el origen a la medio poblacional y normalizando cada coordenada por su desviación estándar. En la Fig. 2.8 mostramos las dos primeras componentes principales de los datos anatómicos. Observamos que mientras que la CP 1 parece separar hombres y mujeres, la CP 2 correlaciona fuertemente con la edad de los participantes.

Con el objetivo de interpretar las direcciones CP 1 y CP 2, clasificamos cada una de sus componentes según cuatro aspectos: lateralidad (izquierda, derecha o central), tipo de medida (volumen, área o espesor), tipo de tejido y ubicación en el cerebro (lóbulo para las estructuras corticales y subcortical para el resto). Para cada aspecto, separamos las componentes en dos conjuntos según su signo, y calculamos la longitud de la proyección de cada CP sobre el subespacio de cada clasificación. Por poner un ejemplo, si consideramos el aspecto “lateralidad”, nos quedamos con aquellas componentes que corresponden a estructuras del hemisferio izquierdo, y separamos las positivas y negativas en dos vectores a los que les calculamos su longitud. En la Fig. 2.9 mostramos los resultados obtenidos, luego de escalear cada eje por la media de la distribución correspondiente a computar las intensidades de direcciones tomadas al azar. Esta normalización es necesaria, ya que en la base original, las distintas características que constituyen los vértices de los polígonos aparecen con distintas frecuencias. Por ejemplo, de las 294 componentes del vector de características anatómicas, hay más componentes lateralizadas en ”izquierdo” (143) o ”derecho” (143) que componentes ”centro” (8). Asimismo, hay más componentes que describen características corticales (190) que de ganglios basales (8).

Vemos que para la primer CP, exceptuando espesores, todas las características tienden a aumentar al movernos a lo largo de la dirección. En consecuencia, concluimos

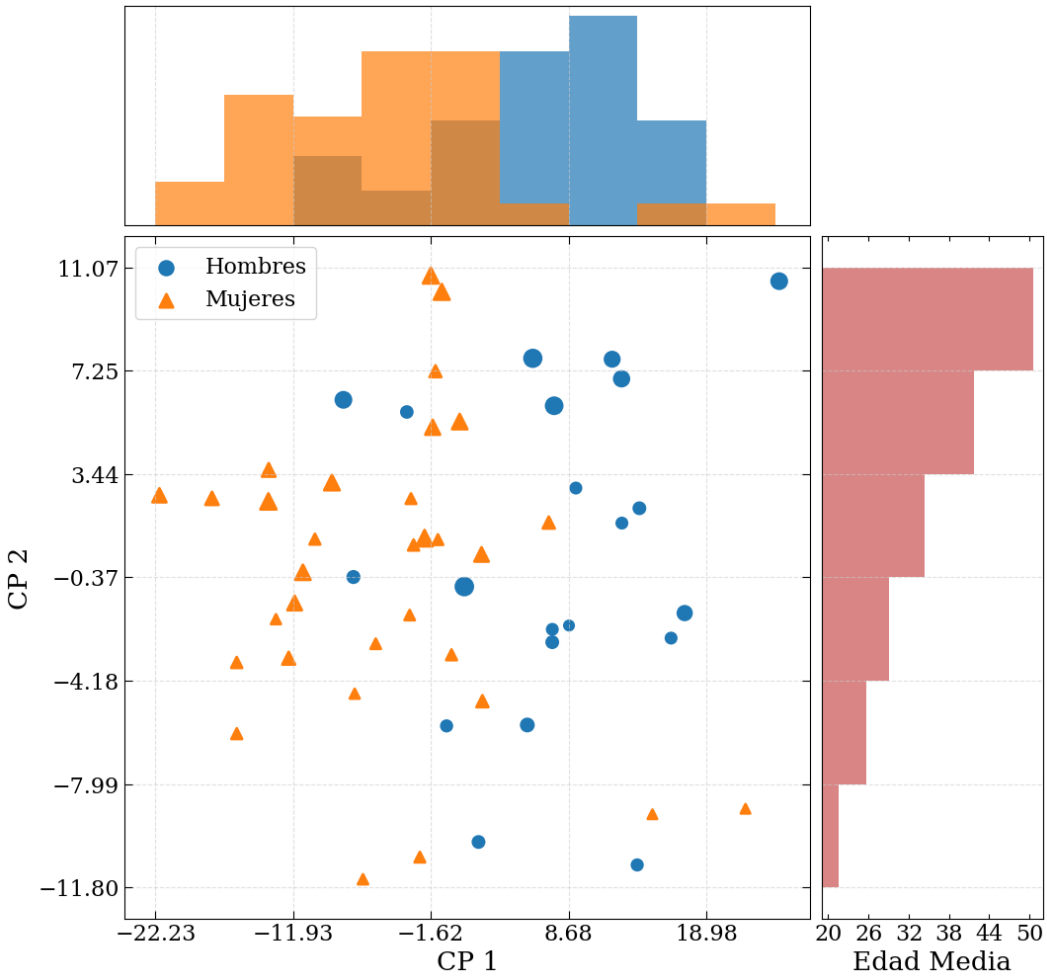

Figura 2.8: Primeras dos componentes principales de los datos anatómicos de la población de San Carlos de Bariloche. Cada punto representa un participante, cuya edad es proporcional al tamaño del punto. Pre-procesmos los datos corriendo el origen a la media poblacional y normalizando cada coordenada por su desviación estándar. En la parte superior mostramos histogramas del número de hombres y mujeres sobre intervalos de la coordenada CP 1. A la derecha mostramos la edad media de los participantes que se encuentran en distintos intervalos de la coordenada CP 2.

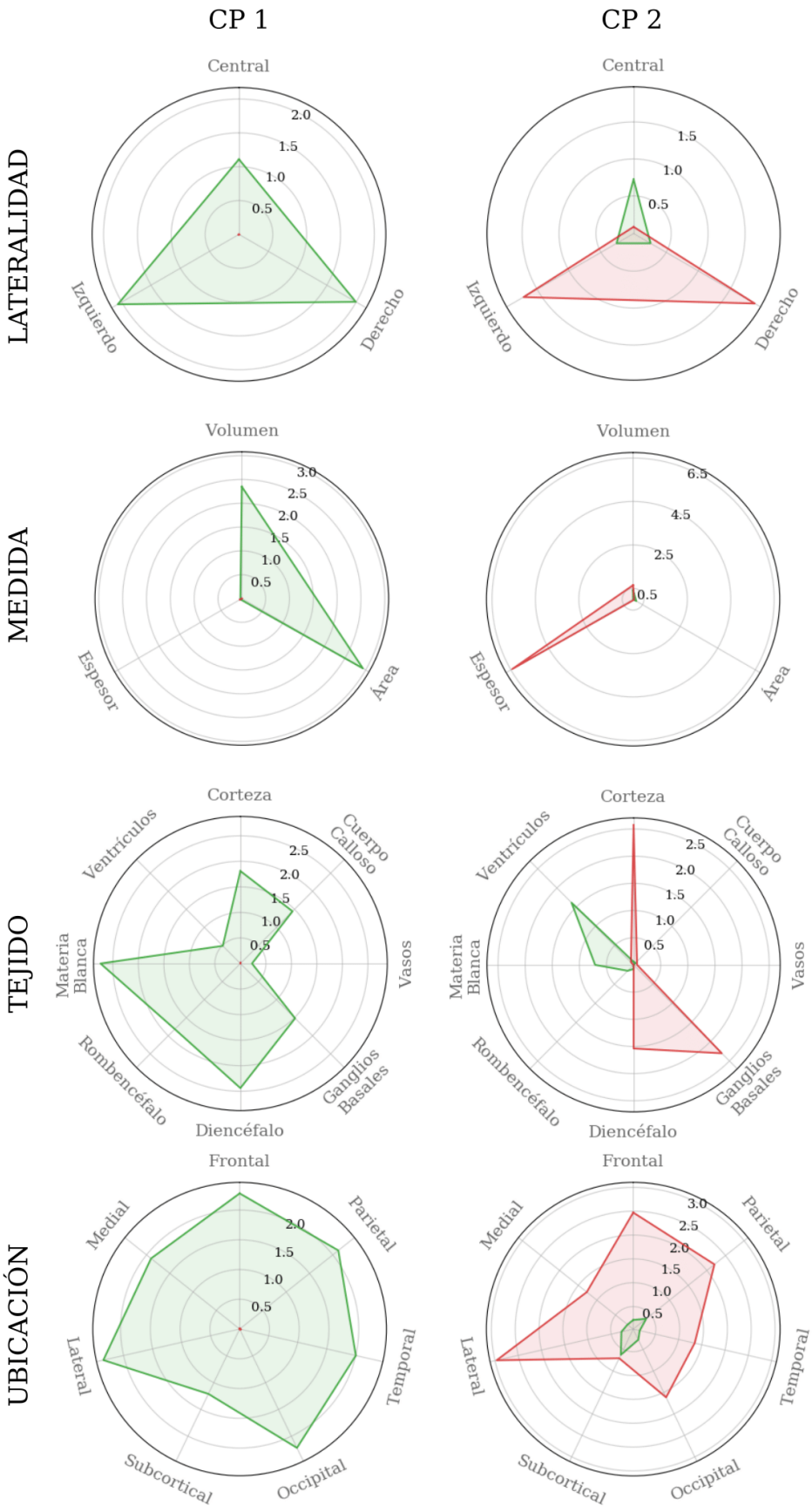

Figura 2.9: Descripción anatómica del significado de las dos componentes principales de la Fig. 2.8. Para cada característica (lateralidad, medida, tejido y ubicación), mostramos la longitud de la proyección de las componentes positivas y negativas sobre cada subespacio en colores verde y rojo respectivamente. Los valores de las proyecciones están normalizados con la media de las distribuciones correspondientes a tomar versores al azar.

que la dirección que maximiza la varianza de los datos y pareciera separar hombres y mujeres, representa una noción de tamaño. Es decir que los hombres tendrían un cerebro de mayor tamaño que las mujeres. Sin embargo, ya exploramos este resultado en la tesis de licenciatura donde mostramos que, más que al sexo, se asocia a una diferencia de tamaño corporal. Es decir, los hombres suelen tener un mayor tamaño corporal que las mujeres, y en particular, también un mayor tamaño cerebral.

Por otro lado, a medida que avanzamos en la dirección de la segunda CP (que a su vez está correlacionada con la edad), vemos que el cerebro comienza a centralizarse, es decir, segmentos correspondientes a hemisferios laterales disminuyen y estructuras centrales aumentan. Por otro lado, mientras que volumen y área permanecen aproximadamente constantes, el espesor disminuye notablemente. Mientras que los ganglios basales y estructuras de corteza disminuyen, la materia blanca y los ventrículos tienden a aumentar. Por último, estructuras correspondientes al lóbulo lateral disminuyen en gran medida.

Uno de los grandes problemas que presenta PCA es que, al ser las variables obtenidas combinaciones lineales de todas las variables originales, resultan difíciles de interpretar. Si bien las dos primeras componentes principales se pueden asociar a tamaños (CP 1) y espesores (CP 2), el resto de las componentes no se pueden asociar a un aspecto o a una región cerebral específica. Además, en mayor o en menor medida, todas las medidas cerebrales participan de cada componente principal. Si proyectamos una dada componente principal en las direcciones de los ejes coordenados, obtenemos la contribución de cada medida cerebral en la componente principal elegida. Las proyecciones grandes (en módulo) son probablemente significativas, es decir, es altamente probable que cambien sólo mínimamente al cambiar la muestra de sujetos. En cambio, las proyecciones pequeñas (en módulo) son probablemente ruido. Por lo tanto, definir a la componente principal como una combinación lineal de *todas* las medidas anatómicas es discutible. Si pudiéramos incluir únicamente a las medidas significativas, podríamos localizar a la componente principal en una o unas pocas regiones cerebrales. Sin embargo, como el espectro de proyecciones decae suavemente, no resulta evidente cómo determinar cuáles proyecciones son significativas y cuáles no. En consecuencia, en los próximos capítulos diseñamos métodos bayesianos para disminuir la dimensionalidad utilizando criterios de significancia y preservando cierto grado de segregación anatómica.

Capítulo 3

Organización del proceso de inferencia

Dado que el tamaño de nuestra muestra es menor que la dimensión del espacio donde viven los datos, elegimos describirlos a través de unas pocas variables latentes, bajo la hipótesis de que se relacionan con las variables originales mediante transformaciones simples. Restringimos la búsqueda a transformaciones simples debido a que un modelo complejo tendría un mayor número de parámetros, y al tener una muestra pequeña, correríamos el riesgo de sobrefitear los datos.

Como mencionamos en el capítulo anterior, PCA es uno de los algoritmos más utilizados para encontrar un conjunto de variables latentes. Tiene la ventaja de que las componentes principales son una transformación lineal de las variables originales, lo cual implica una operación simple. Sin embargo, este método tiene dos inconvenientes. Por un lado, las componentes principales resultan difíciles de interpretar en términos anatómicos. Por otro, típicamente todas las variables originales participan, en mayor o menor medida, de cada componente principal. Sin embargo, no todas las participaciones son significativas. Es decir, al variar la muestra, algunos de los coeficientes de la combinación lineal que define cada componente principal fluctúan fuertemente, particularmente aquellos que son pequeños en módulo.

Una manera de solucionar estos problemas es que, al reducir la dimensión, cada una de las nuevas variables sea combinación lineal de pocas de las originales: sólo de aquellas cuyas correlaciones estén lo suficientemente convalidadas en la muestra para poder asegurar su significancia. Es decir, uno puede desarrollar un criterio para partir el conjunto de variables originales en subconjuntos pequeños compuestos de variables significativamente correlacionadas, que definen variables latentes. De este modo, tales nuevas variables están compuestas por pocas de las originales, permitiendo que sean fácilmente interpretables en términos anatómicos. El problema es entonces cómo elegir una buena partición. Una primera opción es agrupar las variables usando un modelo

conceptual de los datos. Por ejemplo, en nuestro caso los datos representan medidas anatómicas de distintas estructuras cerebrales, así que podríamos agrupar las variables por ubicación espacial, tipo de tejido, funcionalidad, lateralidad, etc. A pesar de que estos criterios se basan en conocimientos previos (lo cual es valioso), a priori no podemos saber si reflejan la estructura estadística de los datos. Es decir, reflejan lo que ya sabíamos de antes, y no la información nueva que resulta de la muestra adquirida.

En el capítulo 4, nosotros optamos por una segunda opción, que consiste en primero inferir la estructura estadística de las variables cerebrales, para luego determinar la partición que mejor la refleje. Para ello, modelamos nuestros datos con una distribución gaussiana, de manera tal que su matriz de precisión (es decir, la inversa de la matriz de covarianza) indica relaciones estadísticas entre pares de variables, y en consecuencia, la estructura que nos interesa. Al inferir tal matriz, tomamos una postura conservadora en la que estimamos únicamente las relaciones significativamente distintas de cero. A continuación, utilizando métodos de teoría de grafos, en el capítulo 5 buscamos la mejor partición en bloques de la matriz, definiendo así subconjuntos cuasi-independientes de variables altamente relacionadas. Una pregunta interesante que abordaremos en el capítulo 5 es si esta partición es en alguna medida similar a la resultante de un modelo conceptual. Una vez determinada la partición óptima, buscamos variables latentes que describan cada uno de los subconjuntos. Para ello, hicimos la hipótesis de que el subespacio generado por cada subconjunto es resultado de un mapeo ruidoso desde un espacio latente de menor dimensión, e inferimos las variables que describen tales espacios latentes. Concretamente, el algoritmo implementado se conoce como PCA Bayesiano, desarrollado en el capítulo 6.

Posteriormente, en el capítulo 7, caracterizamos la influencia de las variables socioambientales sobre las medidas anatómicas del cerebro, suponiendo que tal influencia es mediada por las variables latentes. Siendo más concretos, extendemos el modelo gaussiano permitiendo que la media de la distribución de las variables latentes dependa linealmente de las variables socioambientales. De esta manera, el espacio latente actúa como un cuello de botella de información, en relación a la interacción socioambiental - anatómica. Finalmente, en el capítulo 7 describimos la dependencia entre variables socioambientales y anatómicas, tal como se refleja en nuestra muestra.

Capítulo 4

Inferencia de la estructura estadística de las variables anatómicas

En este capítulo presentamos el modelo utlizado tanto para describir los datos anatómicos observados, como para encontrar la estructura subyacente de las variables medidas.

4.1. Modelo

Representamos los datos anatómicos del i-ésimo participante como un vector aleatorio $\boldsymbol{y}^i \in \mathbb{R}^{d_a}$ que contiene las $d_a = 294$ medidas de volumen, área o espesor de estructuras cerebrales. Dado que las componentes de $\boldsymbol{y}^i$ no son dimensionalmente homogéneas y son cantidades definidas positivas, redefinimos sus coordenadas mediante el mapeo

$$y_k^i \mapsto \ln\left(\frac{y_k^i}{\mathrm{TIV}_i^{m_k}}\right),$$

donde TIV_i es el volumen intracraneal del i-ésimo voluntario y m_k es un coeficiente que vale

$$m_k = \begin{cases} 1, & \text{si } y_k^i \text{ representa un volumen} \\ 2/3, & \text{si } y_k^i \text{ representa un área} \\ 1/3, & \text{si } y_k^i \text{ representa un espesor.} \end{cases}$$

De esta manera, las componentes de $\boldsymbol{y}^i$ pasan a ser cantidades adimensionales. Dado que en la sección 2.2 concluimos que la primera componente principal de PCA representa en buena medida el volumen cerebral total, a partir de ahora normalizamos por el volumen intracraneal, para focalizarnos en efectos independientes de esta medida sobre volúmenes, áreas y espesores de estructuras anatómicas. Con el objetivo de capturar

influencias sobre el tamaño global mediante una única variable, añadimos el volumen intracraneal total como una nueva componente del vector $\boldsymbol{y}$. En lo que sigue reservamos el índice i para participantes, y los índice k y k' para componentes anatómicas.

Dado que el tamaño de nuestra muestra es menor que la dimensión del espacio donde viven los datos, buscamos una representación de la estructura estadística de las medidas anatómicas que refleje relaciones simples. En consecuencia, describimos la dependencia entre las distintas componentes de $\boldsymbol{y}^i$ mediante correlaciones de primer orden, bajo la hipótesis de que los términos de orden superior son menos relevantes, y de ser necesario, podrían capturarse con una expansión a mayor orden. Al no tener información que sustente hipótesis adicionales sobre las correlaciones de orden superior, adoptamos el modelo de mínima estructura [10] consistente con la descipción de correlaciones de primer orden. Minimizar la estructura implica maximizar la entropía [11], y dado un conjunto de restricciones, la forma de la distribución de máxima entropía se puede encontrar fácilmente de manera analítica. Para el caso particular que nos interesa, la distribución de máxima entropía que describe correlaciones de primer orden es

$$\boldsymbol{y}^i \sim \mathcal{N}(\boldsymbol{\mu}, \Omega^{-1}), \tag{4.1}$$

donde parametrizamos la distribución gaussiana mediante la matriz de precisión Ω en lugar de la matriz de covarianza $C = \Omega^{-1}$.

El atractivo de trabajar con la matriz de precisión es que sus elementos indican dependencias directas entre pares de variables. Siendo más precisos, $\Omega_{kk'} = 0$ si y solo si las componentes y^i_k y $y^i_{k'}$ son estadísticamente independientes al condicionar sobre el resto. Entonces, que el elemento $\Omega_{kk'}$ no se anule implica la existencia de una relación estadística entre y^i_k y $y^i_{k'}$ que no es mediada por ninguna otra componente. En cambio, los términos no diagonales de C cuantifican covarianzas entre las componentes respectivas, esté o no tal covarianza mediada por otra variables. En este contexto los elementos de Ω miden relaciones estadísticas directas entre pares de variables. Para ser más precisos, definimos la relevancia de la dependencia directa entre y^i_k y $y^i_{k'}$ como la magnitud del correspondiente elemento de Ω, $|\Omega_{kk'}|$.

4.1.1. Estructura lógica del proceso de inferencia

Un punto crucial para cumplir el objetivo de este trabajo es estimar los elementos de matriz de Ω. Dado que en nuestro caso el número de muestras es menor a la dimensión del espacio donde viven los datos, corremos el riesgo tanto de subestimar como de sobreestimar la relevancia de relaciones estadísticas entre pares de variables. Ciertamente, no podemos solucionar ambos problemas. Por eso, en este trabajo tomamos una postura conservadora, y antes de estimar un elemento de matriz $\Omega_{kk'}$, evaluamos

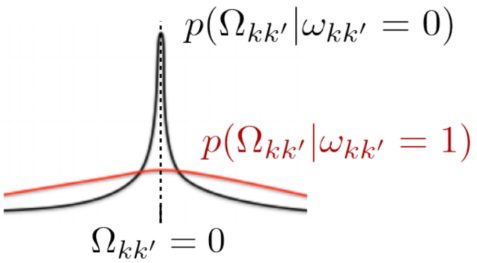

Figura 4.1: Distribuciones a priori de los elementos $\Omega_{kk'}$ para ambos valores de los indicadores latentes $\omega_{kk'}$.

en qué medida los datos brindan evidencia de que existe una dependencia directa entre las estructuras k y k'. La evaluación se realiza introduciendo indicadores latentes binarios $\omega_{kk'} \in \{0, 1\}$, que toman el valor 0 en ausencia de evidencia convincente, y 1 caso contrario. El valor de $\omega_{kk'}$ determina qué prior $p(\Omega_{kk'}|\omega_{kk'})$ se utiliza a continuación para inferir el valor de $\Omega_{kk'}$. Es decir, cuando $\omega_{kk'} = 0$, usamos una prior y cuando $\omega_{kk'} = 1$, usamos otra distinta. Más precisamente, de manera independiente para cada par de índices (k, k'), con $1 \leq k < k' \leq d_a$, la distribución a priori del elemento $\Omega_{kk'}$ es

$$p\left(\Omega_{kk'}|\omega_{kk'}\right) = \left(\frac{\gamma_1}{2}\, e^{-\gamma_1|\Omega_{kk'}|}\right)^{\omega_{kk'}} \left(\frac{\gamma_0}{2}\, e^{-\gamma_0|\Omega_{kk'}|}\right)^{1-\omega_{kk'}}, \tag{4.2}$$

donde $\gamma_1 \ll \gamma_0$ son constantes positivas fijas. Solo especificamos la distribución para elementos por encima de la diagonal debido a que aquellos que se encuentran por debajo quedan determinados por la simetría de Ω. En la Fig. 4.1 mostramos el efecto que tienen las variables binarias $\omega_{kk'}$ sobre la distribución de $\Omega_{kk'}$. Si un indicador binario se anula, la distribución a priori para el correspondiente elemento de Ω se concentra en cero y su valor estimado es pequeño. Por el contrario, si un indicador binario es igual a uno, la distribución a priori se achata y el elemento tiene cierta libertad para tomar valores grandes. En consecuencia, interpretamos $\omega_{kk'} = 1$ como un indicador de que la relación estadística entre y_k^i y $y_{k'}^i$ es relevante.

Para modelar nuestra incerteza sobre las variables binarias $\omega_{kk'}$ suponemos que siguen una distribución de Bernoulli:

$$p\left(\omega_{kk'}|\eta\right) = \begin{cases} \eta, & \text{si } \omega_{kk'} = 1 \\ 1-\eta, & \text{si } \omega_{kk'} = 0, \end{cases} \tag{4.3}$$

donde el parámetro η puede interpretarse como la proporción de pares (k, k') con dependencias directas significativas. Como no tenemos ningún motivo para suponer que η se ubica en alguna región particular del intervalo $[0, 1]$, le asignamos una distribución a priori uniforme:

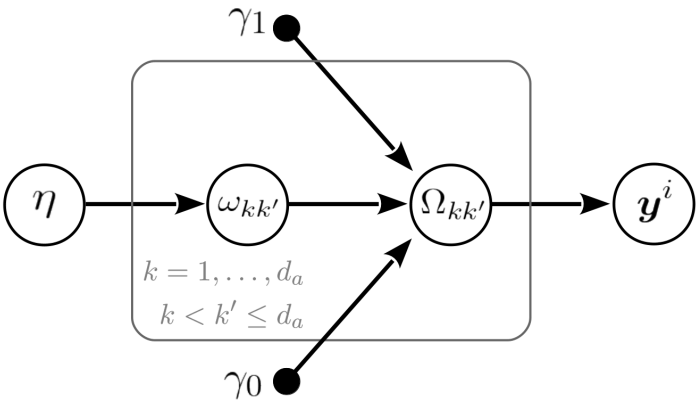

Figura 4.2: Representación gráfica del modelo probabilístico utilizado para describir el proceso generativo de los datos anatómicos observados. Los círculos llenos corresponden a constantes positivas fijas.

$$p(\eta) = \begin{cases} 1, & \text{si } \eta \in [0,1] \\ 0, & \text{si } \eta \notin [0,1]. \end{cases} \tag{4.4}$$

Por último, resta especificar distribuciones a priori para los elementos de la diagonal de Ω. Independientemente para cada índice $k \in \{1, ..., d_a\}$, asignamos la distribución

$$p(\Omega_{kk}) = \begin{cases} \gamma_1 \, e^{-\gamma_1 \Omega_{kk}}, & \text{si } \Omega_{kk} > 0 \\ 0, & \text{si } \Omega_{kk} \leq 0 \end{cases} \tag{4.5}$$

a los elementos Ω_{kk}. Notamos que el requisito de que Ω sea definida positiva introduce dependencias entre los elementos $\Omega_{kk'}$ que no se ven reflejadas en su distribución a priori. Entonces, terminamos de especificar tal distribución restringiendo su soporte al espacio de matrices definidas positivas. En la práctica esto implica restringirse a algoritmos de estimación de parámetros que se limiten a explorar este espacio.

La Fig. 4.2 muestra una representación gráfica del modelo que resume las dependencias introducidas por las Ecs. 4.1 - 4.5. Resumiendo el rol de las variables del modelo:

- η controla el número de elementos de matriz con distribución a priori ancha,
- $\boldsymbol{\omega}$ indica cuáles son los elementos con distribución a priori ancha,
- Ω indica las relaciones estadísticas directas entre pares de medidas anatómicas, y en consecuencia, es la matriz que nos interesa estimar,

y el cociente entre las constantes γ_0 y γ_1 caracterizan cuán distintas son las dos distribuciones a priori empleadas.

Las Ecs. 4.2 - 4.5 determinan la distribución a priori de la matriz de precisión de la distribución de los datos. Para poder completar la inferencia de la gaussiana de la Ec.

4.1, falta especificar la distribución a priori de la media $\boldsymbol{\mu}$ del modelo. Sin embargo, dado que no tenemos ningún motivo para suponer que $\boldsymbol{\mu}$ se encuentra en alguna región particular de $\mathbb{R}^{d_a}$, optamos por tratar a $\boldsymbol{\mu}$ como un parámetro en lugar de como una variable aleatoria. Esto es equivalente a asignarle una distribución a priori (impropia) constante en todo $\mathbb{R}^{d_a}$.

4.2. Estimación de parámetros

Definiendo el conjunto de nuestros datos $\mathcal{D} = \{\boldsymbol{y}^1, \ldots, \boldsymbol{y}^n\}$ donde $n = 193$, la distribución a posteriori de los parámetros del modelo resulta

$$p\left(\boldsymbol{\mu}, \Omega, \boldsymbol{\omega}, \eta | \mathcal{D}\right) \propto \det(\Omega)^{n/2}\, e^{-\frac{1}{2}\sum_i (\boldsymbol{y}^i - \boldsymbol{\mu})^T \Omega (\boldsymbol{y}^i - \boldsymbol{\mu})}\, e^{-\gamma_1 \sum_k \Omega_{kk}} \times \prod_{k<k'} \left[\eta\, \gamma_1\, e^{-\gamma_1 |\Omega_{kk'}|}\right]^{\omega_{kk'}} \left[(1-\eta)\, \gamma_0\, e^{-\gamma_0 |\Omega_{kk'}|}\right]^{1-\omega_{kk'}}, \tag{4.6}$$

donde

- $\boldsymbol{\mu} \in \mathbb{R}^{d_a}$,
- Ω es una matriz definida positiva de tamaño $d_a \times d_a$,
- $\boldsymbol{\omega}$ es un vector binario de $d_a(d_a - 1)/2$ componentes, y
- $\eta \in [0, 1]$.

Cada uno de estos parámetros aporta información relevante sobre la estructura estadística de nuestros datos, por lo que nos interesa estimarlos todos. Podríamos construir estimadores tomando el valor medio o el argumento del máximo de la distribución a posteriori presentada. Sin embargo, dada la complejidad de la Ec. 4.6, tales enfoques no resultan factibles. Por un lado, el cómputo del valor medio requiere calcular integrales en espacios de muy alta dimensión que no se pueden resolver de forma analítica, y los métodos aproximados disponibles suelen ser lentos y computacionalmente costosos. Por otro lado, la presencia del vector binario $\boldsymbol{\omega}$ implica la necesidad de hacer una exploración exhaustiva sobre las $2^{d_a(d_a-1)/2} \sim 10^{12921}$ cadenas binarias posibles para encontrar el máximo de la Ec. 4,6, un número muchísimo mayor al número de bariones en el universo.

Podemos estimar los parámetros $\boldsymbol{\mu}$, Ω y η marginando sobre el vector de indicadores binarios $\boldsymbol{\omega}$, y tomando el argumento del máximo de la distribución resultante. Es decir, usamos los estimadores

$$\left\{\hat{\boldsymbol{\mu}}, \hat{\Omega}, \hat{\eta}\right\} = \underset{\boldsymbol{\mu}, \Omega, \eta}{\text{argmáx}} \left\{p\left(\boldsymbol{\mu}, \Omega, \eta | \mathcal{D}\right)\right\}, \tag{4.7}$$

donde el problema de optimización es fácilmente resoluble por medio del algoritmo EM (*expectation maximization*) [12]. Estos estimadores fueron computados fijando las constantes positivas $\gamma_1 = 1$ y $\gamma_0 = 50$.

4.3. Representación de la estructura subyacente

Representamos la estructura estadística de las variables presentes en nuestros datos mediante un grafo. Los nodos son medidas anatómicas que se encuentran conectadas por la magnitud de sus relaciones estadísticas directas. En concreto, el peso del enlace entre las componentes y_k^i y $y_{k'}^i$ es igual a la magnitud de su correlación parcial $\rho_{kk'}$ [13]:

$$\rho_{kk'} = -\frac{\Omega_{kk'}}{\sqrt{\Omega_{kk}\Omega_{k'k'}}}.$$

De esta manera el grafo refleja las dependencias condicionales de las variables cerebrales.

4.4. Estructura obtenida

La Fig. 4.3 muestra el grafo que resulta de la representación descipta en la sección anterior. Observamos que la mayoría de las conexiones son verdes, y en consecuencia, representan una correlación parcial positiva. Supongamos que tenemos dos personas con un cerebro idéntico salvo, por ejemplo, en el volumen y el área de una dada estructura cerebral. La positividad de la correlación parcial entre tales medidas implica que si una de ellas aumenta al pasar de un cerebro al otro, es probable que la otra medida también lo haga.

Un segundo aspecto a destacar es que todos los enlaces entre áreas y espesores son negativos. Esto es fácilmente apreciable en la Fig. 4.4, donde presentamos la matriz de adyacencia del grafo. Dado que las estructuras corticales tienen el mismo orden dentro de cada tipo de medida, la diagonal **g** de correlaciones parciales negativas en el bloque área - espesor indica que si aumenta la superficie de cualquier estructura cerebral, también disminuye su espesor, fijadas el resto de las medidas anatómicas (en particular, el volumen de tal estructura). Esto es razonable, si aumentamos el área de un cuerpo, manteniendo su volumen constante, y no alteramos fuertemente su geometría, su espesor debe disminuir. Podemos hacer un análisis similar para las diagonales **e** y **f** de elementos positivos en la parte superior de los bloques volumen - área y volumen - espesor. Al aumentar el volumen de una dada estructura cerebral, manteniendo su espesor (área) fijo, su área (espesor) también aumenta.

Por último, la matriz de la Fig. 4.4 refleja la simetría bilateral de las estructurales cerebrales. Si nos concentramos en los bloques volumen - volumen, área - área y espesor

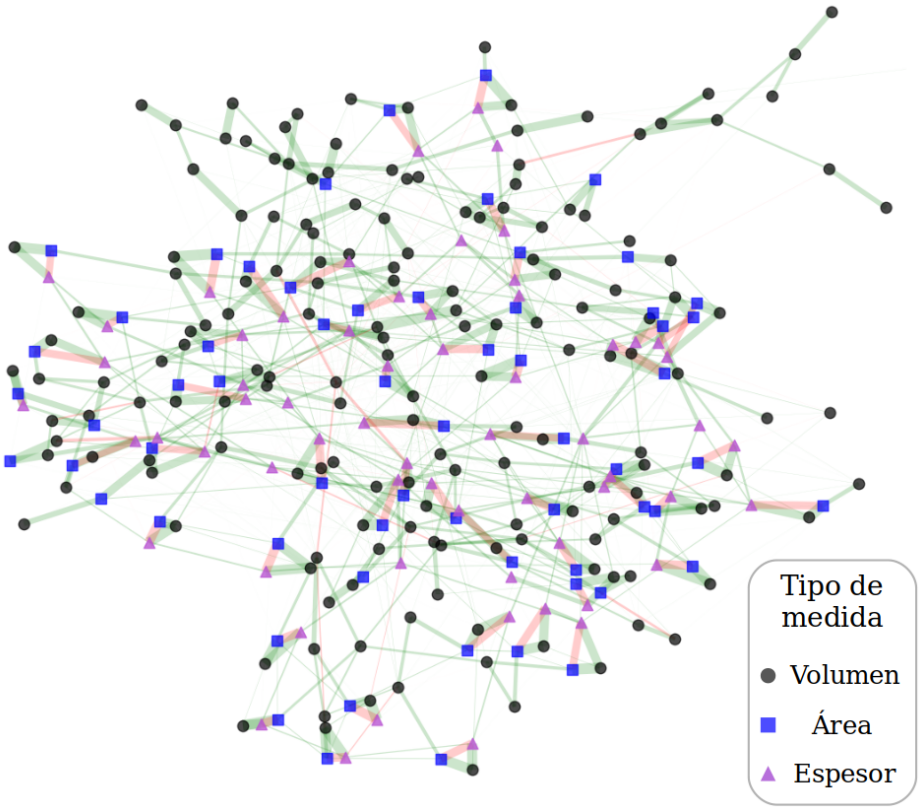

Figura 4.3: Representación en forma de grafo de la estructura de los datos anatómicos. Asignamos misma forma y color a nodos que corresponden a un mismo tipo de medida (volumen, área o espesor). El color (verde o rojo) y espesor de cada enlace indican el signo (positivo o negativo) y la magnitud de la correspondiente correlación parcial.

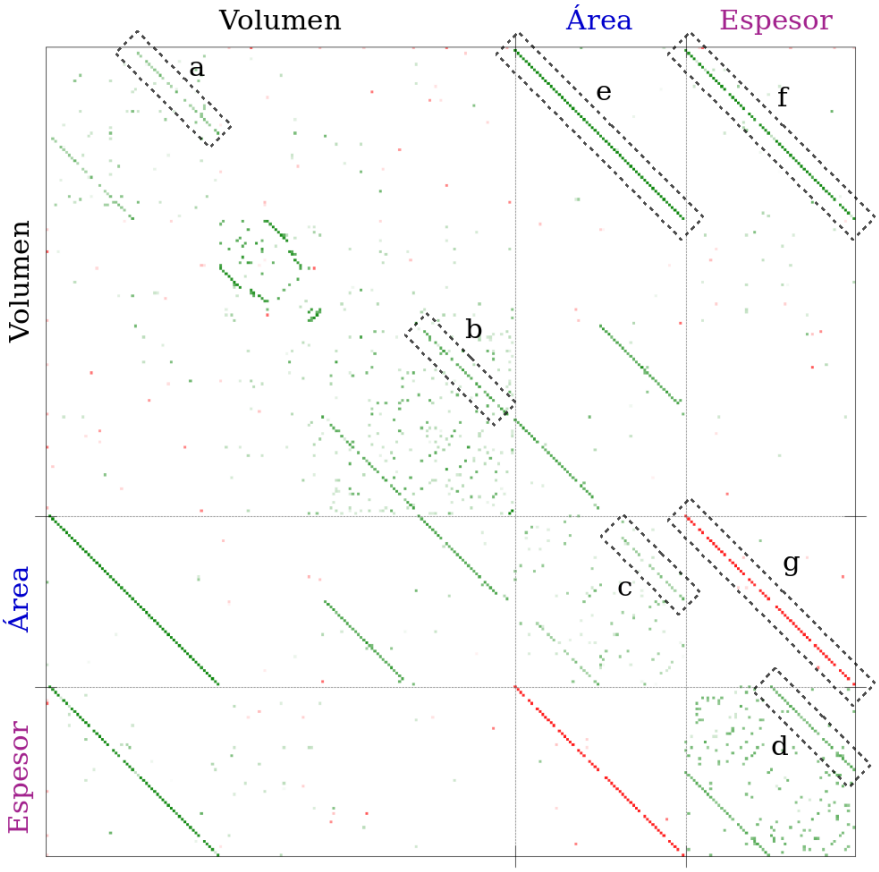

Figura 4.4: Matriz de adyacencia del grafo obtenido. Separamos las medidas por volumen, área y espesor. El color (verde o rojo) y la opacidad de cada elemento indican el signo (positivo o negativo) y la magnitud de la correspondiente correlación parcial. Las diagonales marcadas corresponden a regiones de la matriz discutidas en la sección 4.4.

- espesor, vemos que hay diagonales de elementos positivos. Las diagonales **a-d** corresponden a correlaciones parciales entre las medidas anatómicas de una misma estructura cerebral, pero de hemisferios distintos. Es decir que la mayoría de las medidas de una dada estructura cerebral tienen la siguiente característica: al aumentarlas manteniendo el resto constante, la correspondiente medida del hemisferio opuesto también tenderá a aumentar.

4.5. Dependencia de la estructura inferida con el tamaño de la muestra

Una pregunta interesante es en qué medida depende la estructura presentada en la sección anterior con el tamaño de nuestra muestra. Para ser más concretos, queremos saber si una muestra más grande revelaría aún más relaciones estadísticas relevantes, es decir, ¿necesitamos más datos para inferir la estructura anatómica por completo, o ya estamos en una situación estable?

Para responder esta pregunta, estimamos la estructura estadística de submuestreos de nuestros datos, mediante el mismo procedimiento que datallamos en las secciones 4.1 y 4.2. Para cada tamaño $n' \in \{3, 8, \ldots, 188, 193\}$, tomamos al azar n' de los 193 datos e inferimos el grafo asociado a las medidas anatómicas, repitiendo este proceso 50 veces. La Fig. 4.5 muestra la cantidad de enlaces del grafo inferido en función del tamaño n' de la muestra. Distinguimos dos regiones,

- para $n' \leq 50$ el grafo obtenido no tiene enlaces, y en consecuencia, el modelo determina que todas las variables son independientes. En esta situación, no podemos detectar relaciones estadísticas relevantes, por lo que no es posible hacer inferencia.

- Para $n' > 50$ el grafo comienza a reflejar dependencias entre las medidas anatómicas. En esta región, cuanto más grande sea la muestra, mayor será el número de relaciones relevantes detectadas, y por lo tanto, se vuelve posible inferir cierto grado de estructura de los datos.

Podemos determinar si necesitamos más datos a partir de la estabilidad de la curva graficada. A pesar de que parece estar convergiendo a un número de enlaces ≈ 800, esto podría ser consecuencia del procedimiento utilizado para generar la curva. Dado que estamos tomando submuestreos de un mismo conjunto de datos, a medida que n' se parezca más al tamaño de la muestra total, más dependientes serán los submuestreos. Es decir, cuando $n' \approx 193$, estamos estimando el grafo a partir de prácticamente los mismos datos en todas las realizaciones. Esto necesariamente induce un efecto de convergencia en la curva. Si analizamos la Fig. 4.5 en la región $n' < 100$, de manera

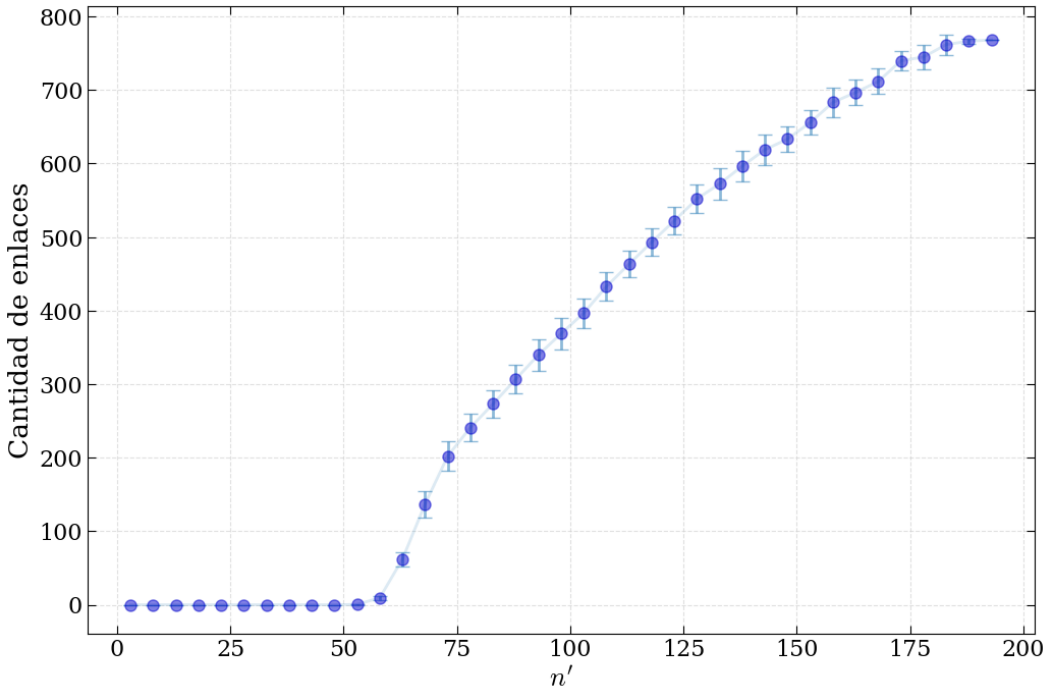

Figura 4.5: Cantidad de enlaces del grafo estimado en función del tamaño n' de la muestra utilizada. Para cada tamaño de la muestra, hicimos 50 submuestreos de los 193 participantes, y graficamos el número medio de enlaces presentes en la estructura inferida. Las barras de error corresponden a $\pm$ una desviación estándar.

tal que los submuestreos sean, en cierta medida, independientes, vemos que la curva no llega a estabilizarse. En consecuencia, concluimos que una muestra más grande nos proporcionaría información útil para inferir más estructura de los datos.

Capítulo 5

Determinación de la partición óptima

En este capítulo mostramos cómo descomponer la matriz de correlaciones parciales en bloques cuasi-independientes. Primero establecemos un mapeo entre este problema y la búsqueda de la estructura de comunidades de un grafo. Luego usamos el algoritmo de Louvain para determinar la partición en comunidades que maximice la modularidad del grafo encontrado en el capítulo anterior.

5.1. Estructura de las correlaciones parciales

En el capítulo anterior inferimos las correlaciones parciales entre todos los pares de variables anatómicas, y representamos esta estructura mediante un grafo. El siguiente paso consiste en usar esta información para definir subconjuntos cuasi-independientes de variables altamente relacionadas. Es decir, buscamos agrupar las medidas cerebrales de manera tal que, al reordenar las componentes del vector anatómico $\boldsymbol{y}^i$ por grupos, la matriz de correlaciones parciales quede cuasi-diagonal a bloques.

La Fig. 5.1 muestra la matriz de correlaciones parciales del grafo anatómico (la misma que vimos en la Fig. 4.4), con las medidas anatómicas convenientemente ordenadas para que la estructura de bloques cuasi-independientes resulte evidente. Si bien todavía no detallamos el procedimiento que permite encontrar la partición en comunidades y el ordenamiento óptimo (el algoritmo se explica en la próxima sección), el resultado es sorprendente: las conexiones fuertes se ubican sobre la diagonal, y los distintos bloques parecen tener siempre la misma estructura interna, o al menos, una muy similar. Por ejemplo, si amplificamos la esquina inferior derecha de la matriz, se hace evidente que los bloques que componen la diagonal tienen todos estructuras internas similares.

A continuación detallamos el algoritmo que permite hacer evidente esta estructura.

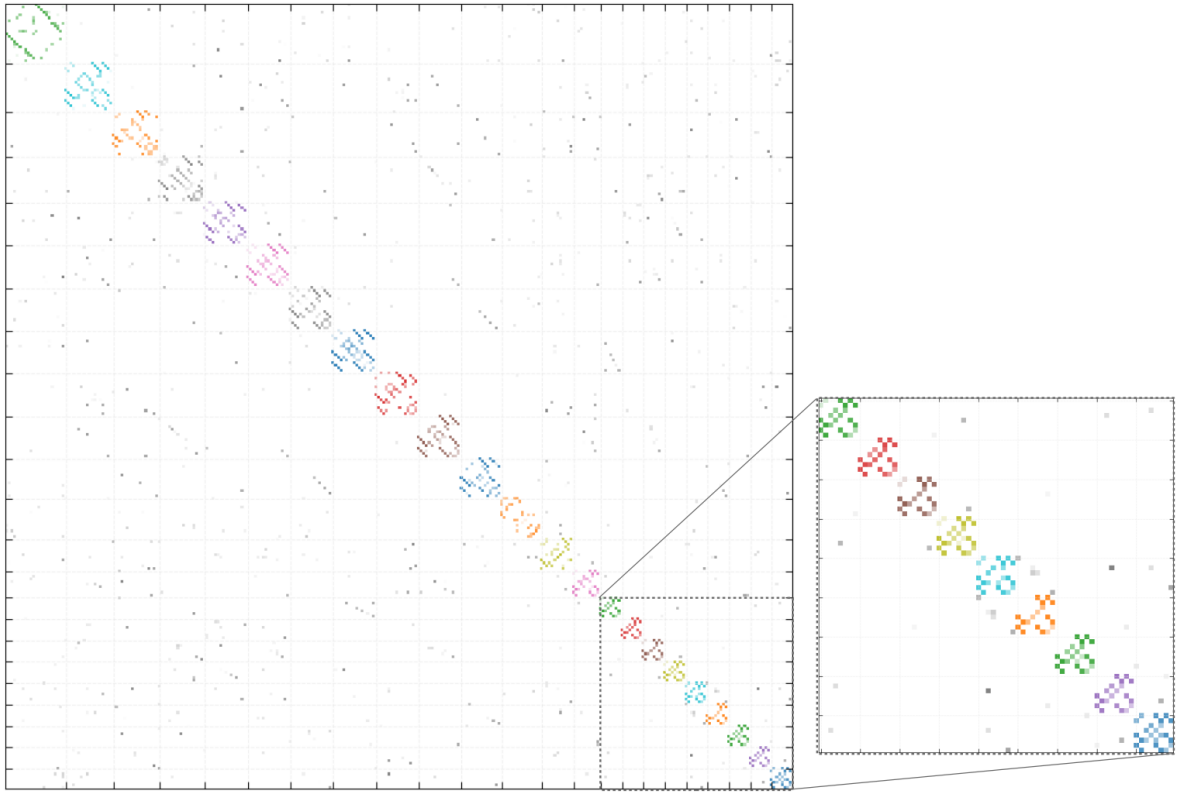

Figura 5.1: Matriz de correlaciones parciales entre pares de variables anatómicas. Mayor opacidad implica una correlación parcial de mayor magnitud. Cada fila y columna corresponde a una componente anatómica, y se encuentran ordenadas según la partición del grafo en comunidades (ver sección 5.2). Elementos en gris representan correlaciones parciales entre componentes de comunidades distintas. Detallamos aquellas comunidades que contienen medidas de una única estructura cerebral.

5.2. Descomposición en bloques cuasi-independientes

Buscamos agrupar las medidas anatómicas en comunidades. Este problema puede formularse en términos de teoría de grafos. Buscar bloques cuasi-independientes es equivalente a buscar subconjuntos de nodos con conexiones internas numerosas y fuertes, y conexiones externas escasas y débiles. En teoría de grafos estos subconjuntos se conocen como comunidades, y el problema de la descomposición en bloques se traduce a detectar comunidades en el grafo anatómico. La Fig. 5.2 muestra un ejemplo de la estructura de comunidades de un grafo.

5.2.1. Algoritmo de Louvain

El algoritmo de Louvain [14] es uno de los métodos más utilizados para resolver el problema de detección de comunidades. Primero se asocia al grafo una cantidad denominada *modularidad* que, dada una partición en comunidades, mide la densidad de conexiones intra-comunidad en relación a la densidad media de conexiones que tal partición presentaría en un grafo con conexiones al azar. El algoritmo busca la partición en comunidades que maximiza la modularidad, que se define como

$$Q = \frac{1}{2m} \sum_{k,k'} \left(|\rho_{kk'}| - \frac{s_k s_{k'}}{2m} \right) \delta_{c_k, c_{k'}},$$

donde $|\rho_{kk'}|$ es el peso del enlace entre los nodos k y k', s_k es la suma de los pesos de los enlaces que involucran al nodo k, m es la suma del peso de todos los enlaces del grafo, c_k es la comunidad del nodo k, δ es la delta de Kronecker, y los índices k y k' recorren todos los nodos del grafo. El algoritmo de Louvain consiste en buscar la partición en comunidades que resulte en el máximo Q, determinando de forma automática el número óptimo de comunidades.

El proceso de optimización involucra dos etapas que se repiten alternadamente. En la primera etapa cada nodo comienza en su propia comunidad, y secuencialmente se los

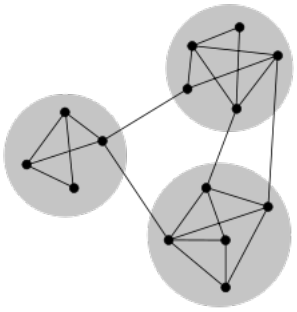

Figura 5.2: Esquema de un grafo con su respectiva estructura de comunidades. En este ejemplo hay tres subconjuntos de nodos que internamente están densamente conectados, y cuyas conexiones con otros subconjuntos son débiles.

asigna a la comunidad que, en cada paso, logre el mayor incremento de la modularidad del grafo, hasta alcanzar un máximo local. La segunda etapa del algoritmo agrupa los nodos pertenecientes a una misma comunidad y construye un nuevo grafo cuyos nodos son las comunidades de la etapa previa. A continuación se aplica la primera etapa al nuevo grafo, repitiendo este proceso hasta alcanzar un máximo de modularidad.

5.3. Comunidades neuroanatómicas encontradas

Utilizando el algoritmo de Louvain, determinamos la partición en comunidades que maximiza la modularidad del grafo anatómico obtenido en el capítulo anterior. La Fig. 5.3 muestra las comunidades obtenidas. Los colores de los nodos son los mismos que los de la matriz de correlaciones parciales de la Fig. 5.2.

Las comunidades encontradas son: una comunidad con regiones ocupadas por líqudo encefalorraquídeo y materia blanca con conexiones de largo alcance, como el cuerpo calloso y el quiasma óptico, una comunidad con regiones subcorticales y corticales filogenéticamente antiguas, y 21 comunidades corticales, en su mayoría conteniendo estructuras homólogas de ambos hemisferios cerebrales y funcionalmente conectadas. De estas 23 comunidades, 9 contienen medidas anatómicas que corresponden a una única estructura cerebral, con medidas que pertenecen a tanto el hemisferio izquierdo como el derecho. La Fig. 5.4 ejemplifica este caso con las comunidades (a), (b) y (c) marcadas en la Fig. 5.3. Otras 12 comunidades contienen medidas de exactamente dos estructuras cerebrales bilaterales. La Fig. 5.5 muestra el ejemplo (d) de la Fig. 5.3. Por último, 2 comunidades mezclan medidas de numerosas estructuras. La Fig. 5.6 describe el ejemplo marcado como (e) en la Fig. 5.3. Las subsecciones que siguen describen estos ejemplos, demostrando que el algoritmo de segregación en comunidades obtiene resultados que son razonables desde el punto de vista anatómico, funcional y evolutivo.

5.3.1. Comunidades de una única región cerebral

Las comunidades que corresponden a una única región cerebral presentan una estructura interna extremadamente similar, como se evidencia en el detalle de la Fig. 5.1. En todos los casos, la medida de volumen está positivamente correlacionada con tanto el espesor como el área de la región en cuestión. El espesor y el área, sin embargo, están negativamente correlacionados entre sí. La parte de la comunidad que se ubica en el hemisferio izquierdo se conecta con la homóloga del hemisferio derecho exclusivamente a través de los espesores y del volumen de materia blanca, en ambos casos, con una correlación positiva. En consecuencia, los espesores y el volumen de materia blanca a uno y otro lado del cerebro son más parecidos entre sí que los volúmenes o las áreas.

La Fig. 5.4(a) muestra la comunidad formada por las medidas de la porción orbital

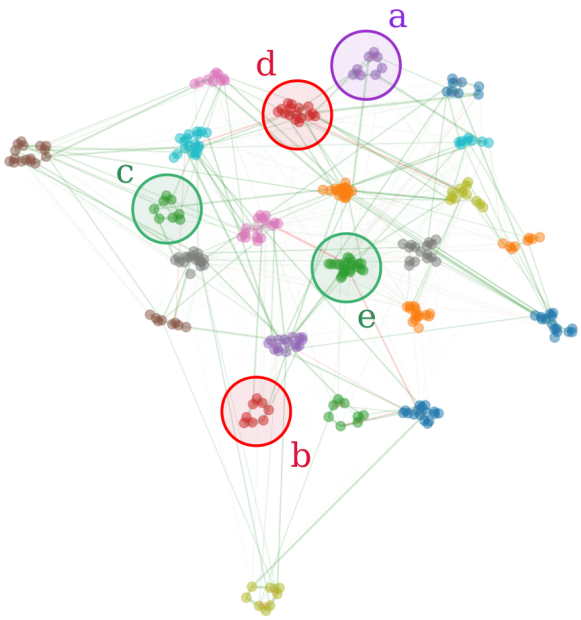

Figura 5.3: Estructura de comunidades del grado de componentes anatómicas. El color de cada conexión (verde o roja) indica el signo de la correspondiente correlación parcial (positiva o negativa respectivamente). El espesor de cada conexión indica la magnitud de la correlación parcial. Se marcan cinco comunidades que se describen en detalle en las Figs. 5.4-5.6

.

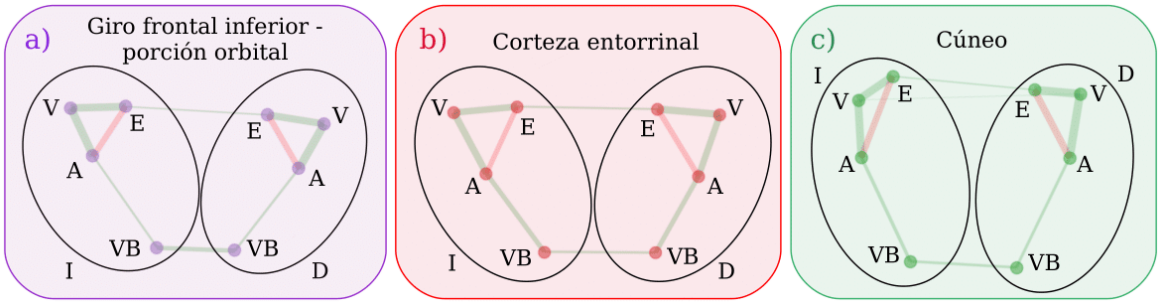

Figura 5.4: Estructura interna de tres comunidades que contienen medidas de una única región cerebral: **a)** porción orbital del giro frontal inferior, **b)** corteza entorrinal, y **c)** cúneo. Las tres comunidades se muestran en la Fig. 5.3, con la misma convención de colores. D: hemisferio derecho, I: hemisferio izquierdo, V: volumen de materia gris, VB: volumen de materia blanca, A: área, E: espesor.

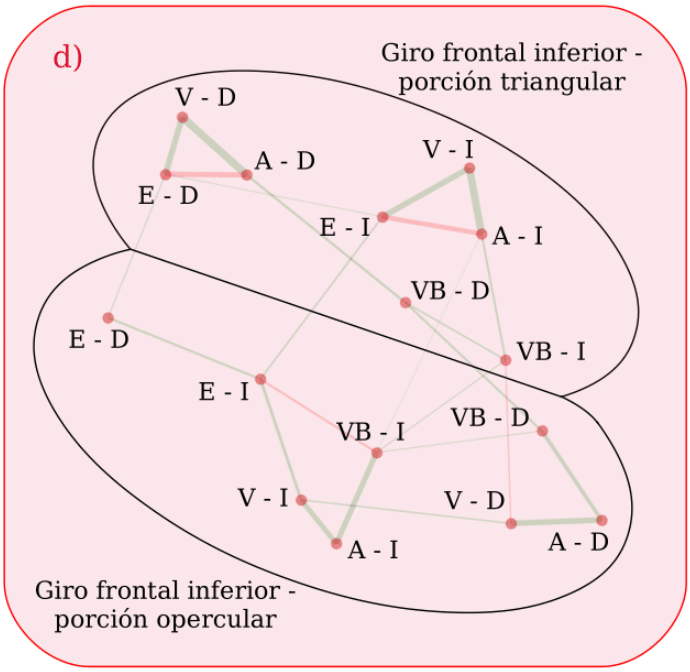

Figura 5.5: Estructura interna de un ejemplo de comunidad que contienen medidas de dos regiones cerebrales: la comunidad **(d)**, con el área de Broca del lado izquierdo, y la homóloga del lado derecho. Convenciones de líneas, colores y siglas como en la Fig. 5.4.

del giro frontal inferior. Anatómicamente, esta estructura está sumamente relacionada con el área de Broca, que es un complejo con varias estructuras cerebrales asociadas al lenguaje, ubicadas exclusivamente en el hemisferio dominante, típicamente el izquierdo. El algoritmo de clustering, sin embargo, le asignó una comunidad propia, separada del resto de las regiones del lenguaje del área de Broca, y preservando su estructura bilateral. Esto puede estar relacionado con una diferencia funcional, ya que hay estudios que muestran que, a diferencia de las estructuras del área de Broca, la porción orbital del giro frontal está involucrada en la compresión semántica [15, 16], y con el reconocimiento de emociones [17].

En las Figs. 5.3(c) y 5.3(d) mostramos las comunidades que contiene las medidas de la corteza entorrinal y el cúneo respectivamente. La corteza entorrinal suele asociarse a tareas de memoria, navegación y percepción del tiempo [18–20], mientras que el cúneo está involucrado en el procesamiento visual y el control inhibitorio en pacientes con depresión bipolar [21, 22]. Enfatizamos una vez más que las comunidades (a), (b) y (c) de la Fig. 5.4 tienen una estructura interna que es prácticamente idéntica, a pesar de que están involucradas en funciones cognitivas marcadamente distintas.

5.3.2. Ejemplo de una comunidad con 2 regiones cerebrales

La comunidad (d) que mostramos en la Fig. 5.5 contiene todas las medidas de las porciones triangular y opercular del giro frontal inferior, de ambos hemisferios. Estas

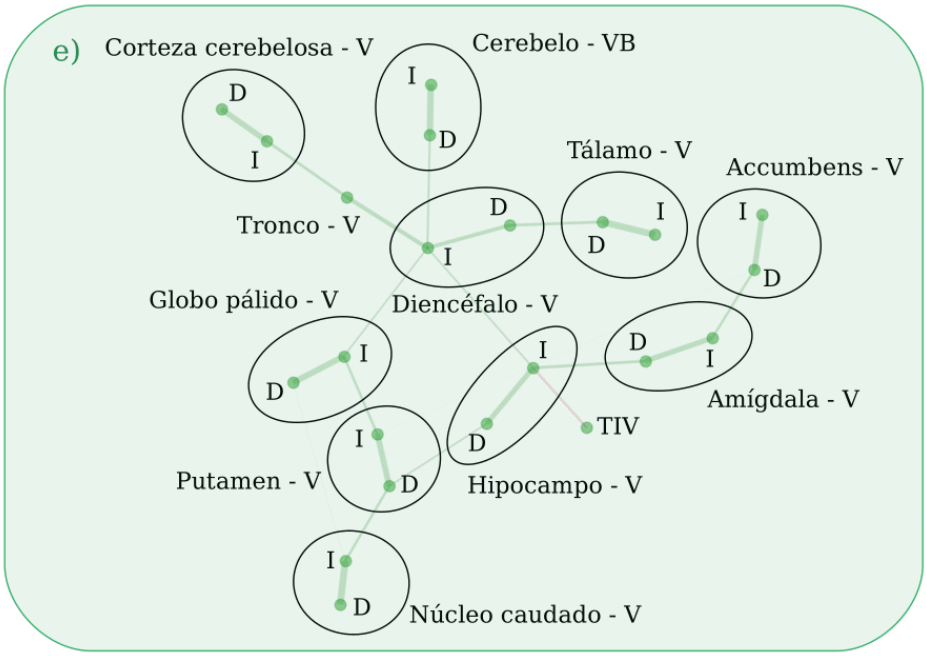

Figura 5.6: Estructura interna de un ejemplo de comunidad que contienen medidas filogenéticamente antiguas, y corresponde a la comunidad **(e)** de la Fig. 5.3. Convenciones de líneas, colores y siglas como en la Fig. 5.3.

dos regiones son anatómicamente aledañas. En el hemisferio dominante, conforman el área de Broca y están principalmente involucradas en la producción y comprensión del lenguaje [23–25]. En el hemisferio no dominante, participan mayormente en la ejecución de tareas motoras [26, 27]. La comunidad contiene la porción triangular de ambos hemisferios, y la estructura de conexiones es similar a la observada en las comunidades de una única estructura cerebral (Fig. 5.4, comunidades (a-c)). Sin embargo, la porción opercular es asimétrica, probablemente reflejando la especialización mayormente unilateral de algunas funciones lingüísticas.

5.3.3. Ejemplo de comunidad mixta

La comunidad mixta (e) que mostramos en la Fig. 5.6 está compuesta por medidas de estructuras filogenéticamente antiguas. La mayor parte son estructuras subcorticales, salvo por aquellas correspondientes al hipocampo, que si bien es parte de la corteza cerebral, se diferencia del resto de la corteza por tener una estructura de capas más sencilla, y más primitiva desde el punto de vista evolutivo. Llama la atención que las comunidades obtenidas por el algoritmo de clustering agrupan estructuras que surgieron en etapas definidas de la diferenciación de especies. En particular, esta estructura es perfectamente bilateral en aquellas medidas que se expresan bilateralmente. Contiene medidas del cerebelo, que suele asociarse a funciones motoras, lenguaje, atención, y algunos rasgos emocionales y de comportamiento [28–32]; medidas de los ganglios basales (putamen, núcleo caudado, globo pálido, accumbens), que se relacionan con ta-

reas motoras, motivación, toma de decisiones y memoria [33–36]; medidas del sistema límbico (hipocampo, amígdala), que regula emociones, comportamiento, aprendizaje y memoria [37–39]; y medidas del diencéfalo, que funciona como una estación de relevo en la entrada y salida de información sensorial y motora [40].

Capítulo 6

Búsqueda del espacio latente

Para cada una de las comunidades encontradas en el capítulo anterior, determinamos el conjunto de variables latentes que mejor describe el subespacio generado por las componentes anatómicas involucradas, bajo la hipótesis de que las nuevas variables son resultado de un mapeo lineal de las originales. Esto conllevó la implementación de un proceso de inferencia que especifica de forma automática el número de variables latentes necesarias para describir cada comunidad, logrando así una reducción de la dimensión de nuestros datos.

En los dos capítulos anteriores nos concentramos en describir las relaciones estadísticas entre pares de variables, en particular aquellas que no están mediadas por las restantes. Esto nos permitió partir el conjunto original de medidas anatómicas en subconjuntos que presentan interacciones débiles con el resto. Sin embargo, ahora nos interesa explicar la variabilidad de los datos dentro del subespacio asociado a cada comunidad con menos variables. Es decir, buscamos una descripción simplificada que tenga en cuenta toda la estructura estadística de una dada comunidad. Dado que ahora necesitamos una descripción global de cada subespacio, pasamos nuestro foco de la matriz de precisión Ω a la matriz de covarianza C de la distribución sobre las medidas anatómicas.

Podríamos buscar el conjunto de variables latentes aplicando PCA sobre las componentes anatómicas de cada comunidad. Dado que cada comunidad está compuesta por pocas componentes, la nuevas variables serían fácilmente interpretables. Sin embargo, PCA no tiene incorporado un criterio para determinar el número de variables de la descripción simplificada. Por esta razón, utilizamos la extensión del modelo probabilístico de PCA denominada PCA bayesiano [41], donde se determina la dimensionalidad efectiva de cada espacio latente de manera automática como parte del proceso de inferencia bayesiana.

6.1. PCA bayesiano

6.1.1. Formulación probabilística de PCA

La formulación probabilística de PCA describe los datos observados a partir del siguiente proceso generativo [42] adaptado a nuestro problema particular. Para cada comunidad c, se introduce el vector $\boldsymbol{z}_c^i \in \mathbb{R}^{d_{l_c}}$ de las d_{l_c} medidas latentes del i-ésimo voluntario, cuya distribución de probabilidad a priori es la gaussiana

$$p(\boldsymbol{z}_c^i) = \mathcal{N}(\mathbf{0}, \mathbb{1}).$$

Las d_{a_c} medidas observadas $\boldsymbol{y}_c^i \in \mathbb{R}^{d_{a_c}}$ [1]de la c-ésima comunidad se definen entonces como una transformación lineal de $\boldsymbol{z}_c^i$ con ruido aditivo gaussiano

$$\boldsymbol{y}_c^i = W_c \boldsymbol{z}_c^i + \boldsymbol{\mu}_c + \boldsymbol{\epsilon}_c, \text{ [2]} \tag{6.1}$$

donde W_c es una matriz de tamaño $d_{a_c} \times d_{l_c}$, $\boldsymbol{\mu}_c$ es un vector d_{a_c}-dimensional y $\boldsymbol{\epsilon}_c$ es un vector aleatorio que sigue la distribución

$$p(\boldsymbol{\epsilon}_c) = \mathcal{N}(\mathbf{0}, \sigma_c^2 \mathbb{1}).$$

Por lo tanto,

$$p(\boldsymbol{y}_c^i | \boldsymbol{z}_c^i) = \mathcal{N}(W_c \boldsymbol{z}_c^i + \boldsymbol{\mu}_c, \sigma_c^2 \mathbb{1}).$$

La distribución marginal de las variables observadas queda entonces dada por

$$p(\boldsymbol{y}_c^i) = \int_{\mathbb{R}^{d_{l_c}}} \mathrm{d}\boldsymbol{z}_c \; p(\boldsymbol{y}_c^i | \boldsymbol{z}_c) p(\boldsymbol{z}_c) = \mathcal{N}(\boldsymbol{\mu}_c, C_c), \tag{6.2}$$

donde la matriz de covarianza de la c-ésima comunidad es $C_c = W_c W_c^t + \sigma_c^2 \mathbb{1}$.

El modelo dado por la Ec. 6.2 representa una distribución gaussiana gobernada por los parámetros $\boldsymbol{\mu}_c$, W_c y σ_c. Los autores de [42] mostraron que el estimador de *maximum likelihood* de W_c es

$$\hat{W}_c = U_c (\Lambda_c - \sigma_c^2 \mathbb{1})^{1/2},$$

donde Λ_c es una matriz diagonal con los d_{l_c} autovalores más grandes de la matriz de

[1] Recordamos que al comienzo de la sección 4.1 redefinimos $\boldsymbol{y}^i$ de manera tal que sus componentes pasen a ser cantidades adimensionales.

[2] Dada la forma de la Ec. 6.1, la elección particular para la media y matriz de covarianza de la distribución sobre $\boldsymbol{z}_c^i$ no supone pérdida de generalidad.

covarianza muestreada S_c a partir de las componentes de la comunidad c

$$S_c = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_c^i - \bar{\boldsymbol{y}}_c)(\boldsymbol{y}_c^i - \bar{\boldsymbol{y}}_c)^t,$$

siendo

$$\bar{\boldsymbol{y}}_c = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_c^i,$$

y U_c es la matriz cuyas columnas son los d_{l_c} autovectores correspondientes. Es decir que la solución de este modelo probabilístico describe al espacio latente como aquel que retiene la máxima varianza, donde toda variabilidad adicional de los datos es explicada por el ruido de fondo $\boldsymbol{\epsilon}_c$. Este resultado permite interpretar el modelo generativo dado por la Ec. 6.1 como una formulación probabilística de PCA.

6.1.2. Extensión bayesiana

En [41] se extendió esta formulación para incorporar una manera de determinar la dimensión efectiva d_{l_c} del espacio latente de cada comunidad como parte del proceso de inferencia bayesiana. Consiste en introducir una distribución a priori $p(\boldsymbol{\mu}_c, W_c, \sigma_c)$ sobre los parámetros del modelo, para luego computar la distribución a posteriori $p(\boldsymbol{\mu}_c, W_c, \sigma_c|\{\boldsymbol{y}_c^1, \ldots, \boldsymbol{y}_c^n\})$, a partir de la cual es posible estimar los valores de los parámetros.

Al igual que en la sección 4.1.1, nosotros optamos por no asignar una distribución a priori sobre la media $\boldsymbol{\mu}_c$ del modelo, ya que no tenemos ningún motivo para suponer que se encuentra en alguna región particular de $\mathbb{R}^{d_{a_c}}$.

En el caso de cada W_c, se introduce una distribución jerárquica $p(W_c|\boldsymbol{\beta}_c)$, gobernada por el vector d_{l_c}-dimensional de hiperparámetros $\boldsymbol{\beta}_c = (\beta_{c,1}, \ldots, \beta_{c,d_{l_c}})$. Cada hiperparámetro controla una de las columnas de W_c mediante una distribución gaussiana de la forma

$$p(W_c|\boldsymbol{\beta}_c) = \prod_{m=1}^{d_{l_c}}\left(\frac{\beta_{c,m}}{2\pi}\right)^{d_{l_c}/2}\exp\left(-\frac{\beta_{c,m}}{2}||\boldsymbol{w}_{c,m}||^2\right),$$

donde $\boldsymbol{w}_{c,m}$ es la m-ésima columna de W_c. Cada $\beta_{c,m}$ controla la precisión del correspondiente $\boldsymbol{w}_{c,m}$, de forma tal que si un dado $\beta_{c,m}$ tiene una distribución a posteriori concentrada en valores grandes, $\boldsymbol{w}_{c,m}$ tenderá a tener componentes nulas, y la correspondiente coordenada $z_{c,m}$ del espacio latente no tendrá influencia sobre las variables observadas. En este sentido, una columna de W_c con elementos iguales a cero implica una dirección del espacio latente que es irrelevante para explicar la variabilidad de los datos. En consecuencia, se fija la dimensión del espacio latente de la c-ésima comunidad en su máximo valor posible d_{a_c}, y se define su dimensión efectiva d_{l_c} como el número de columnas de W_c cuya magnitud es distinta de cero.

Completamos el modelo bayesiano definiendo una distribución a priori sobre las componentes de $\boldsymbol{\beta}_c$. Concretamente, hacemos la siguiente elección

$$p(\boldsymbol{\beta}_c) = \prod_{m=1}^{d_{a_c}} \frac{b^a \, \beta_{c,m}^{a-1} \, e^{-b\beta_{c,m}}}{\Gamma(a)},$$

donde fijamos las constantes $a = b = 10^{-3}$ para obtener distribuciones a priori anchas.

6.1.3. Estimación del ruido

A diferencia de la extensión bayesiana original de PCA, nosotros no asignamos una distribución a priori a la magnitud σ_c del ruido de fondo $\boldsymbol{\epsilon_c}$. En lugar de introducir otra distribución al modelo, estimamos el ruido de fondo con las mediciones repetidas que presentamos en la sección 2.1.5. Es decir, consideramos que $\boldsymbol{\epsilon}_c$ es la variabilidad intrasujeto, resultado de errores tanto en la adquisición de la resonancia como en la segmentación de la imagen. Bajo la hipótesis de que cada etapa de la adquisición de datos introduce ruido aditivo que no depende de etapas previas, el teorema central del límite asegura que podemos aproximar la distribución de $\boldsymbol{\epsilon}_c$ mediante la gaussiana

$$p(\boldsymbol{\epsilon}_c) = \mathcal{N}(\mathbf{0}, \Psi_c),^{3}$$

donde estimamos la matriz de covarianza Ψ_c de las medidas de la comunidad c con las mediciones repetidas. Con esta modificación, la matriz de covarianza que aparece en la Ec. 6.2 resulta $C_c = W_c W_c^t + \Psi_c$.

6.1.4. Estimación de parámetros

Dado que las columnas de W_c indican la dimensión del espacio latente de la c-ésima comunidad, y sus filas permiten interpretar la acción que tiene cada variable latente sobre las medidas anatómicas, resulta crucial estimar sus elementos para simplificar nuestra descripción de los datos. Al mismo tiempo, nos interesa capturar la influencia de las variables socioambientales a través de la descripción simplificada, y por lo tanto, también necesitamos estimar las medidas latentes $Z_c \equiv \{\boldsymbol{z}_c^1, \ldots, \boldsymbol{z}_c^n\}$ de todos los voluntarios.

En teoría, dada una comunidad c, el modelo formulado permite calcular la distribución $p(Z_c, W_c, \boldsymbol{\beta}_c | \{\boldsymbol{y}_c^1, \ldots, \boldsymbol{y}_c^n\})$ a partir de la cual podemos construir estimadores. Sin embargo, dada la complejidad de tal distribución, en la práctica no resulta factible. Por este motivo implementamos un tratamiento variacional que resulta computacionalmente eficiente y permite aproximar la distribución a posteriori de los parámetros

[3] Dada la forma de la Ec. 6.1, podemos fijar la media $\langle \boldsymbol{\epsilon}_c \rangle = \mathbf{0}$ sin pérdida de generalidad.

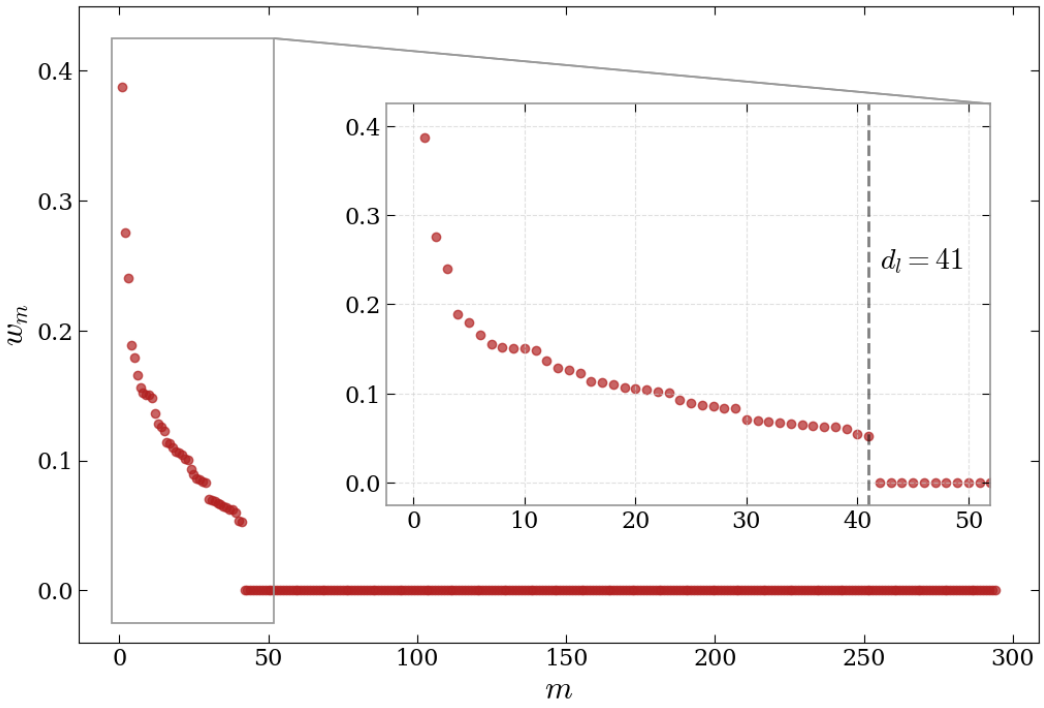

Figura 6.1: Magnitud w_m de las columnas de todas las matrices W_c en orden decreciente. Mostramos en detalle aquellos puntos correspondientes a $w_m > 0$.

del modelo [41].

La inferencia variacional consiste en buscar la distribución $q(Z_c, W_c, \boldsymbol{\beta}_c)$ que sea lo más similar posible a la distribución a posteriori sobre los parámetros del modelo, y simultáneamente cumpla ciertas restricciones que aseguren un algoritmo de estimación eficiente. La noción de disimilitud que utilizamos, y en consecuencia el funcional a minimizar, es la divergencia de Kullback-Leibler

$$D_{KL}\left(q \,||\, p\right) = \int \mathrm{d}Z_c \,\mathrm{d}W_c \,\mathrm{d}\boldsymbol{\beta}_c \; q(Z_c, W_c, \boldsymbol{\beta}_c) \,\ln\left[\frac{q(Z_c, W_c, \boldsymbol{\beta}_c)}{p(Z_c, W_c, \boldsymbol{\beta}_c | \{\boldsymbol{y}_c^1, \ldots, \boldsymbol{y}_c^n\})}\right]. \quad (6.3)$$

Al mismo tiempo, aseguramos que el algoritmo que permite computar q sea tratable exigiendo que se factorice como

$$q(Z_c, W_c, \boldsymbol{\beta}_c) = q(Z_c)\, q(W_c)\, q(\boldsymbol{\beta}_c). \quad (6.4)$$

Entonces el cómputo de q consiste en minimizar el funcional de la Ec. 6.3 sujeto a la restricción dada por la Ec. 6.4. Una vez obtenida la distribución q, estimamos los

parámetros del modelo asociados a la c-ésima comunidad como

$$\hat{Z}_c = \int \mathrm{d}Z_c \; q(Z_c)\, Z_c = \langle Z_c \rangle_q,$$
$$\hat{W}_c = \int \mathrm{d}W_c \; q(W_c)\, W_c = \langle W_c \rangle_q,$$
$$\hat{\boldsymbol{\beta}}_c = \int \mathrm{d}\boldsymbol{\beta}_c \; q(\boldsymbol{\beta}_c)\, \boldsymbol{\beta}_c = \langle \boldsymbol{\beta}_c \rangle_q.$$

6.2. Parámetros estimados

En la Fig. 6.1 mostramos, en orden decreciente, la magnitud $||\boldsymbol{w}_m||$ de las columnas de todas las matrices W_c estimadas, donde se ve claramente un salto entre las columnas nulas y las no nulas. Dado que solo hay 41 columnas con componentes no todas nulas, definimos la dimensión del espacio latente total como

$$d_l = \sum_c d_{l_c} = 41.$$

Como consecuencia de haber tratado cada comunidad por separado, cada una de las componentes del espacio latente afecta a una única comunidad, permitiendo que tales variables sean fácilmente interpretables. En la tabla 6.1 mostramos la dimensión del espacio latente inferido para cada comunidad.

Comunidad	Número de variables latentes
Giro temporal superior Giros temporales transversos (V, A, E) Polo temporal	4
Lóbulo parietal superior Precúneo	4
Giro supramarginal Giro postcentral	4
Cerebelo Ganglios basales Sistema límbico Diencéfalo Volumen intracraneal	3
Giro frontal superior Giro frontal medio (parte caudal) - D Polo frontal	3
Giro temporal medio Giro temporal inferior	3
Giro cingulado posterior Istmo del giro cingulado	3
Área de Broca	3
Ventrículos Plexos coroideos Líquido cefalorraquídeo Cuerpo Calloso Quiasma óptico	3
Giro frontal medio (parte rostral) Giro frontal medio (parte caudal) - I	3
Giro parahipocampal Giro fusiforme	2
Corteza pericalcarina Giro lingual	1
Giro cingulado anterior	1
Giro precentral Giros temporales transversos (VB)	1
Cúneo	1
Lóbulo parietal inferior	1
Giro occipital lateral	1

Tabla 6.1: Número de variables latentes estimadas para cada comunidad. Aquellas comunidades para las que se estimó un espacio latente nulo no se encuentran en la tabla. D: hemisferio derecho, I: hemisferio izquierdo, V: volumen de materia gris, VB: volumen de materia blanca A: área, E: espesor. Si todas las medidas de una dada estructura están presentes en la comunidad, no aclaramos tipo de medida o lateralidad.

Capítulo 7

Influencia de las variables socioambientales

En este capítulo extendemos el modelo presentado en la sección 4.1 con el objetivo de caracterizar la influencia de las variables socioambientales sobre el espacio latente de cada comunidad. Al mismo tiempo, permitimos que el modelo recupere las interacciones anatómicas del tipo intercomunidad y mencionamos las más relevantes.

7.1. Extensión del modelo

Ahora extendemos el modelo de la sección 4.1 para considerar que los participantes tienen datos socioambientales además de los anatómicos. Al mismo tiempo, describimos las características anatómicas de manera simplificada, es decir, a través de las variables latentes estimadas en el capítulo anterior. En concreto, representamos los datos del i-ésimo participante como un par $(\boldsymbol{z}^i, \boldsymbol{x}^i)$ donde $\boldsymbol{z}^i \in \mathbb{R}^{d_l}$ contiene las d_l medidas anatómicas latentes y $\boldsymbol{x}^i \in \mathbb{R}^{d_s}$ contiene las siguientes $d_s = 11$ respuestas de su encuesta socioambiental (ver apéndice A):

- edad E $\in \{18, 19, \ldots, 60\}$,
- sexo S $\in \{\text{femenino} : -1, \text{masculino} : 1\}$,
- altura A,
- peso P,
- mano dominante MD $\in \{\text{izquierda} : -1, \text{ambas} : 0, \text{derecha} : 1\}$,
- peso al nacer PN,
- meses al nacer MN,

- amamantado AM $\in \{\text{no} : 0, \text{sí} : 1\}$,
- número de embarazos EM,
- nivel de educación NE $\in \{\text{P(I)} : 1, \text{P(C)} : 2, \text{S(I)} : 3, \text{S(C)} : 4, \text{U(I)} : 5, \text{U(C)} : 6\}$ (P: primario, S: secundario, U: universitario/terciario, I: incompleto, C: completo),
- y número de idiomas NI.

Con el objetivo de recuperar las interacciones directas entre comunidades más relevantes, permitimos que las componentes de $\boldsymbol{z}^i$ covaríen. Entonces, al igual que en el capítulo anterior, asignamos a $\boldsymbol{z}^i$ una distribución gaussiana

$$\boldsymbol{z}^i \sim \mathcal{N}(\boldsymbol{\mu}, \Omega^{-1}),$$

donde ahora permitimos que $\boldsymbol{\mu}$ y Ω sean distintas del vector nulo y la matriz identidad respectivamente. Más aún, bajo la hipótesis de que términos de orden superior son menos relevantes, introducimos una dependencia lineal de $\boldsymbol{\mu}$ con los factores socioambientales, es decir,

$$\boldsymbol{\mu} = \boldsymbol{\mu}_0 + A\boldsymbol{x}^i,$$

y en consecuencia,

$$\boldsymbol{z}^i \sim \mathcal{N}(\boldsymbol{\mu}_0 + A\boldsymbol{x}^i, \Omega^{-1}),$$

donde los coeficientes de A describen correlaciones entre variables socioambientales y latentes, Ω contiene las correlaciones entre variables latentes y tomamos a $\boldsymbol{\mu}_0$ como un parámetro del modelo al que no le asignamos una distribución a priori.

Con tal de no interpretar correlaciones inexistentes (es decir, con tal de evitar los falsos positivos), estamos dispuestos a tener un cierto número de falsos negativos, y por ende, a renunciar a detectar efectos pequeños. Esta estrategia implica que las correlaciones entre variables anatómicas y socioambientales que detectemos, serán con alta probabilidad relevantes. Lo mismo vale para las correlaciones entre variables latentes. Asignamos entonces distribuciones a priori para los coeficientes de matriz de Ω y de A iguales a las usadas en la sección 4.1. Es decir, en el caso de las interacciones socioambiental-anatómica también distinguimos dos situaciones: una en la que logramos detectar una relación relevante y otra en la que no. Frente a la detección estimamos el correspondiente elemento de A con una distribución a priori ancha y, en caso contrario, con una distribución a priori concentrada en cero.

Formalizamos esta idea introduciendo indicadores latentes $\alpha_{kj} \in \{0, 1\}$ tales que, de manera independiente para cada par de índices (k, j) [1]con $1 \leq k \leq d_l$ y $1 \leq j \leq d_s$,

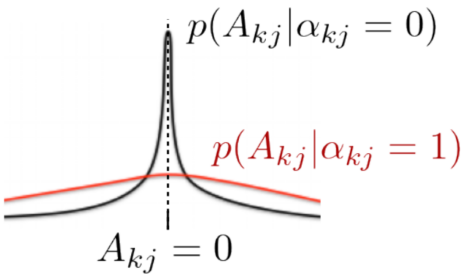

Figura 7.1: Distribuciones a priori de los elementos A_{kj} para ambos valores de los indicadores latentes α_{kj}.

la distribución a priori del elemento A_{kj} sea

$$p\left(A_{kj}|\alpha_{kj}\right) = \left(\frac{\gamma_1}{2}e^{-\gamma_1|A_{kj}|}\right)^{\alpha_{kj}}\left(\frac{\gamma_0}{2}e^{-\gamma_0|A_{kj}|}\right)^{1-\alpha_{kj}}, \tag{7.1}$$

donde γ_1 y γ_0 son las mismas constantes positivas fijas que aparecen en la estructura de distribuciones a priori de Ω. En la Fig. 7.1 el mostramos el efecto de los indicadores latentes sobre la distribución a priori de los elementos de A. De forma análoga a los indicadores binarios $\omega_{kk'}$ de Ω, interpretamos $\alpha_{kj} = 1$ como un indicador de que la componente socioambiental x_j^i tiene un efecto relevante sobre la variable anatómica latente z_k^i.

Completamos el modelo asignando a cada α_{kj} una estructura de distribuciones a priori similar a la correspondiente a los indicadores $\omega_{kk'}$. Concretamente, elegimos

$$p\left(\alpha_{kj}|\theta\right) = \begin{cases} \theta, & \text{si } \alpha_{kj} = 1 \\ 1-\theta, & \text{si } \alpha_{kj} = 0, \end{cases} \tag{7.2}$$

y

$$p\left(\theta\right) = \begin{cases} 1, & \text{si } \theta \in [0,1] \\ 0, & \text{si } \theta \notin [0,1]. \end{cases} \tag{7.3}$$

Mencionando brevemente el rol de cada parámetro:

- θ controla el número de elementos de A con distribución a priori ancha,
- $\boldsymbol{\alpha}$ indica cuáles son los elementos con distribución a priori ancha,
- A indica las relaciones estadísticas del tipo socioambiental-anatómica, y en consecuencia, es una matriz que nos interesa estimar,

y el cociente entre las constantes γ_0 y γ_1 caracterizan cuán distintas son las dos distribuciones a priori empleadas.

[1] Así como en la sección 4.1 reservamos los índices i y k para participantes y componentes anatómicas respectivamente, en lo que sigue nos referimos a variables socioambientales mediante el índice j.

7.2. Estimación de parámetros

Dado que al momento de presentar este trabajo tenemos procesadas únicamente las encuestas socioambientales de la muestra de San Carlos de Bariloche, estimamos los parámetros del modelo sólo con estos datos. Redefiniendo entonces el conjunto de nuestros datos como $\mathcal{D} = \{(\boldsymbol{x}^1, \boldsymbol{z}^1), \ldots, (\boldsymbol{x}^n, \boldsymbol{z}^n)\}$ donde ahora $n = 77$, la distribución a posteriori sobre los parámetros del modelo extendido es tal que

$$\begin{aligned} p\left(A, \boldsymbol{\alpha}, \theta, \Omega, \boldsymbol{\omega}, \eta | \mathcal{D}\right) \propto \det(\Omega)^{n/2}\, e^{-\frac{1}{2}\sum_i (\boldsymbol{z}^i - \boldsymbol{\mu}_0 - A\boldsymbol{x}^i)^T \Omega (\boldsymbol{z}^i - \boldsymbol{\mu}_0 - A\boldsymbol{x}^i)}\, e^{-\gamma_1 \sum_k \Omega_{kk}} \\ \times \prod_{k,j} \left[\theta\, \gamma_1\, e^{-\gamma_1 |A_{kj}|}\right]^{\alpha_{kj}} \left[(1-\theta)\, \gamma_0\, e^{-\gamma_0 |A_{kj}|}\right]^{1-\alpha_{kj}} \\ \times \prod_{k<k'} \left[\eta\, \gamma_1\, e^{-\gamma_1 |\Omega_{kk'}|}\right]^{\omega_{kk'}} \left[(1-\eta)\, \gamma_0\, e^{-\gamma_0 |\Omega_{kk'}|}\right]^{1-\omega_{kk'}}, \end{aligned} \tag{7.4}$$

a partir de la cual podemos inferir las matrices A y Ω.

Una manera de reducir el problema de inferencia a subproblemas conocidos, es marginar sobre los vectores de indicadores binarios $\boldsymbol{\alpha}$ y $\boldsymbol{\omega}$ para luego estimar los parámetros restantes tomando el argumento del máximo de la distribución resultante. Es decir, estimar los parámetros del modelo mediante

$$\left\{\hat{\mathrm{A}}_{\mathrm{mSSL}}, \hat{\theta}_{\mathrm{mSSL}}, \hat{\Omega}_{\mathrm{mSSL}}, \hat{\eta}_{\mathrm{mSSL}}\right\} = \underset{\mathrm{A},\theta,\Omega,\eta}{\operatorname{argmáx}} \left\{p\left(A, \theta, \Omega, \eta | \mathcal{D}\right)\right\}, \tag{7.5}$$

donde el problema de optimización puede resolverse por medio del algoritmo ECM (*expectation conditional maximization*) [43]. Estos estimadores fueron computados fijando $\gamma_1 = 1$.

La Ec. 7.5 ofrece una posible implementación de la idea general de introducir indicadores binarios que detecten relaciones estadísticas relevantes, permitiendo tomar una política conservadora a la hora de hacer inferencias. El detalle de esta implementación en particular puede encontrarse en Deshpande et al. (2019), donde además está detallado el proceso de optimización utilizado para resolver la Ec. 7.5. El procedimiento fue nombrado *multivariate spike-and-slab LASSO* (mSSL), razón por la cual nos referimos a los estimadores de la Ec. 7.5 con el subíndice mSSL [44].

7.3. Parámetros estimados

En la Fig. 7.2 mostramos la representación en forma de grafo de los parámetros estimados, para distintos valores de γ_0. Enlaces entre componentes z_k^i — $z_{k'}^i$ y z_k^i — x_j^i se describen con la magnitud de la correlación parcial $\rho_{kk'}$ y del elemento A_{kj}, respectivamente. A medida que aumentamos γ_0, el grafo estimado pierde conexiones. Esto se debe a que, al aumentar γ_0, la distribución a priori concentrada en cero aumenta

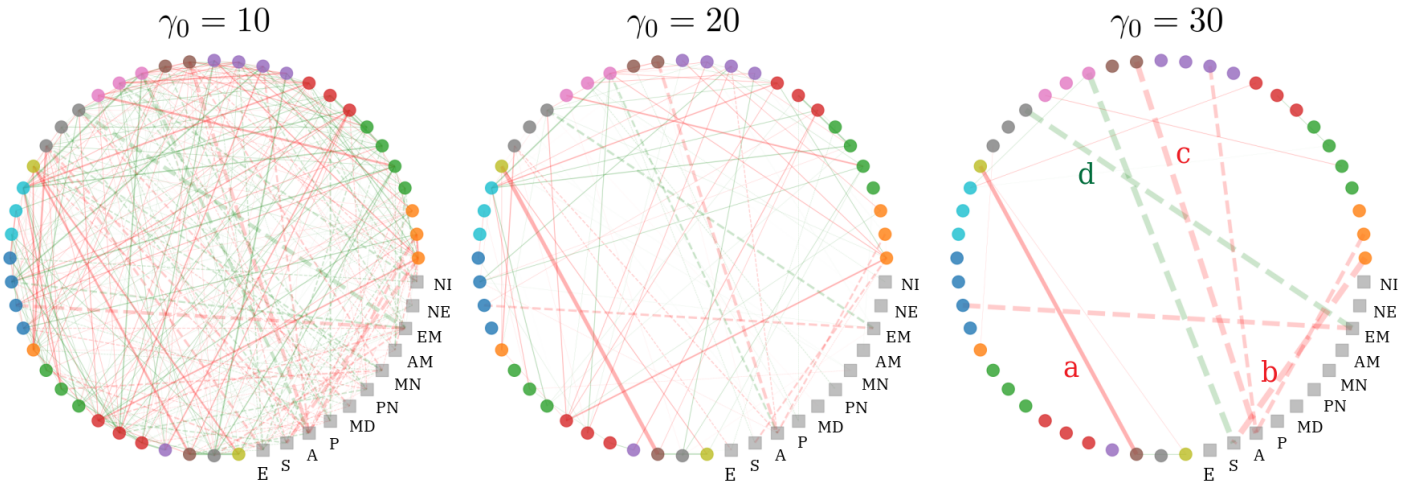

Figura 7.2: Grafo estimado para distintos valores del parámetro γ_0. Círculos adyacentes de un mismo color representan variables latentes de una misma comunidad. Los cuadrados grises corresponden a las variables socioambientales consideradas. Conexiones de trazo continuo y discontinuo representan relaciones del tipo anatómica-anatómica y socioambiental-anatómica respectivamente. El color de cada enlace (verde o rojo) indica el signo (positivo o negativo) de la correspondiente correlación parcial (anatómica-anatómica) o coeficiente de la regresión (socioambiental-anatómica). El espesor del enlace indica la magnitud de la relación. A medida que γ_0 crece, se pierden gradualmente conexiones, pero no se crean nuevas, ni se observan cambios de signo. E: edad, S: sexo, A: altura, P: peso, MD: mano dominante, PN: peso al nacer, MN: meses al nacer, AM: amamantado, EM: número de embarazos, NE: nivel de educación, NI: número de idiomas.

su máximo, y en consecuencia, se vuelve más conveniente fijar $\omega_{kk'} = 0$ y $\alpha_{kj} = 0$. En otras palabras, cuanto más grande sea γ_0, más relevante tiene que ser la dependencia explicada por el correspondiente elemento de A o Ω para que se estime distinto de cero. Por ejemplo, para $\gamma_0 = 30$ el grafo presenta muy pocas conexiones, donde nombramos **a** a la conexión anatómica más relevante, y **b**, **c** y **d** a conexiones que involucran a las variables socioambientales que tienen mayor efecto: sexo, altura y número de embarazos, respectivamente.

7.3.1. Interacción intercomunidad

La Fig. 7.3 muestra la composición anatómica de las comunidades a las que pertenecen las variables latentes involucradas en el enlace **a** de la Fig. 7.2. Notamos que estructuras de una misma comunidad son espacialmente cercanas. Al mismo tiempo, las comunidades involucradas en la interacción son vecinas.

7.3.2. Influencia socioambiental

En el grafo de la Fig. 7.2 asociado a $\gamma_0 = 30$ observamos que las interacciones del tipo socioambiental-anatómica más relevantes involucran al sexo (S), la altura (A) y el número de embarazos (EM). El enlace **b** representa la influencia del sexo sobre una dirección del espacio de medidas anatómicas cuya componente más relevante es el

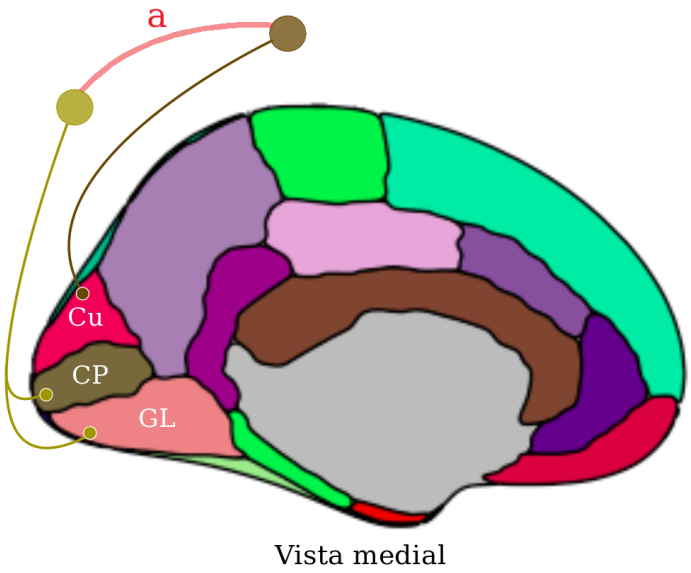

Figura 7.3: Ubicación anatómica de las estructuras presentes en las variables del enlace **a** de la Fig. 7.2. CP: corteza pericalcarina, GL: giro lingual, Cu: cúneo.

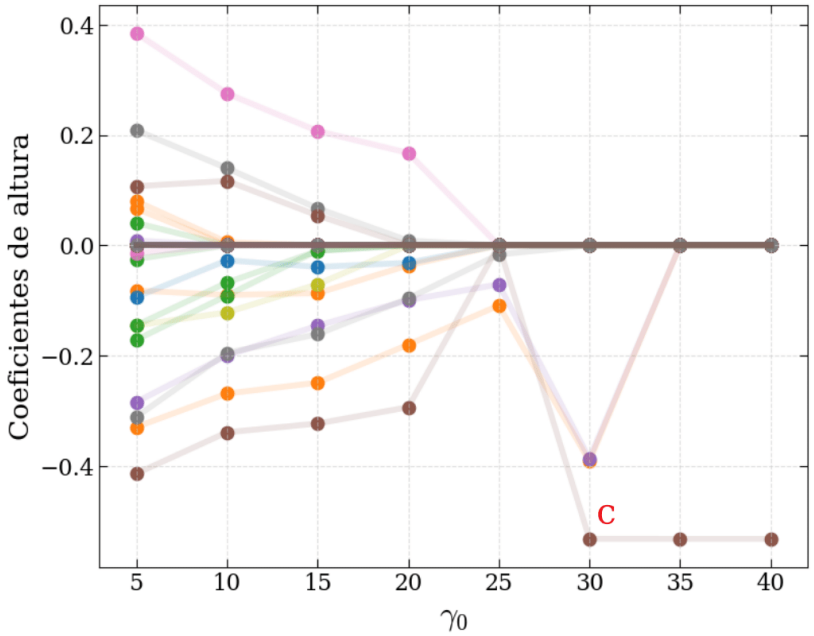

Figura 7.4: Coeficientes de A que representan la influencia lineal de la altura sobre las variables latentes en función del parámetro γ_0. Cada conjunto de puntos unidos por una curva de trazo continuo corresponde a una misma variable latente. Las unidades de los coeficientes son una desviación estándar de la correspondiente variable latente sobre una desviación estándar de la altura. Marcamos el coeficiente correspondiente al enlace **c** de la Fig. 7.2.

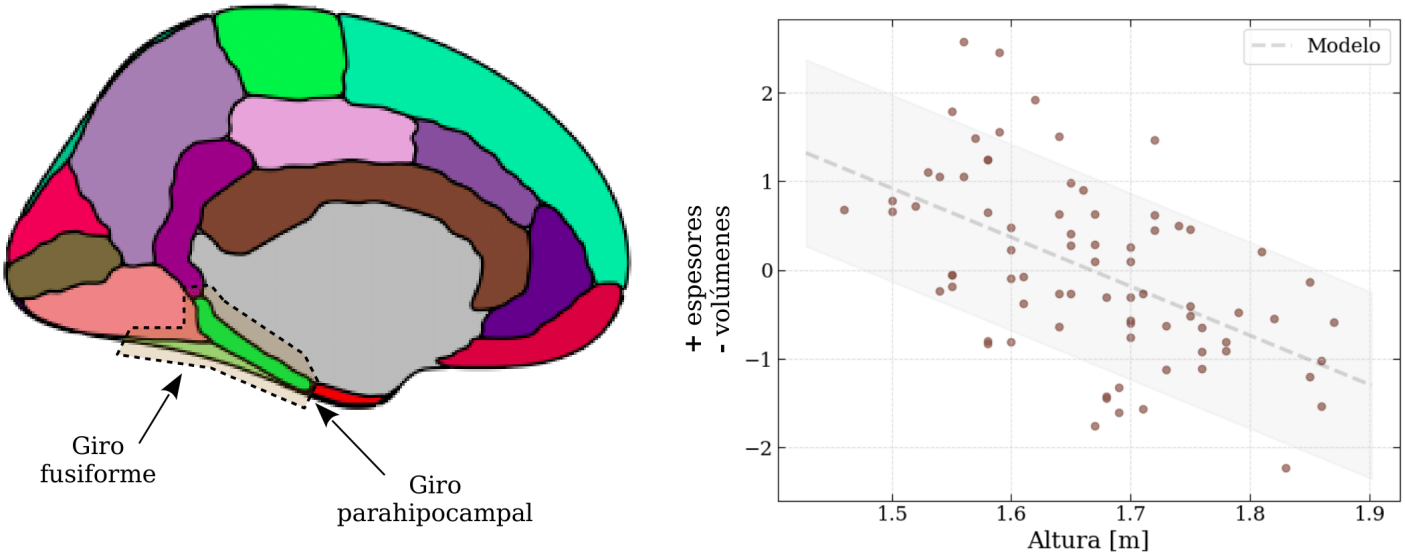

Figura 7.5: Estructuras pertenecientes a la comunidad de la variable latente del enlace **c** de la Fig. 7.2 (izquierda). Coordenada de cada sujeto sobre la dirección anatómica de la variable latente del enlace **c** (en desviaciones estándar) en función de la altura (derecha). Mostramos la dependencia estimada por el modelo como una línea de trazo discontinuo con una franja gris que corresponde a la media ± una desviación estándar de la coordenada.

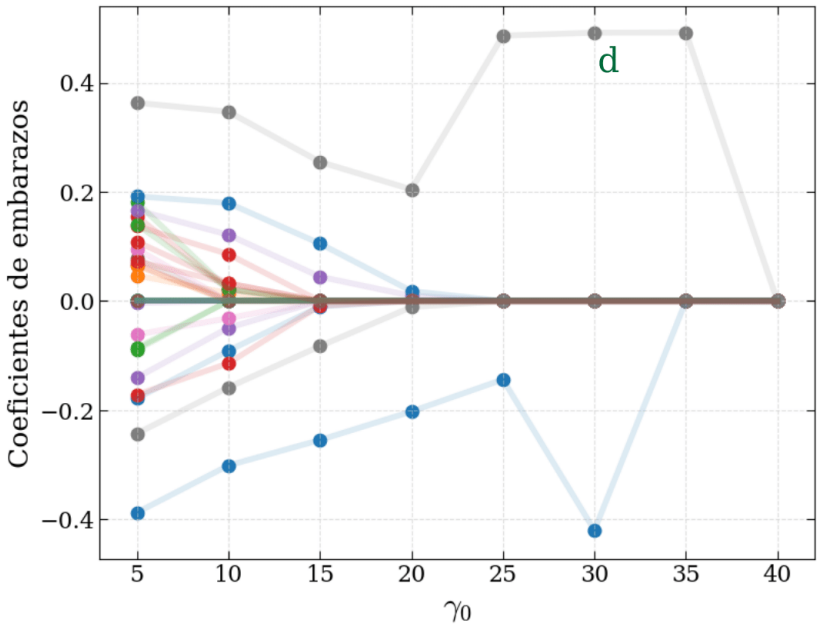

Figura 7.6: Coeficientes de A que representan la influencia lineal del número de embarazos sobre las variables latentes en función del parámetro γ_0. Cada conjunto de puntos unidos por una curva de trazo continuo corresponde a una misma variable latente. Las unidades de los coeficientes son una desviación estándar de la correspondiente variable latente sobre una desviación estándar del número de embarazos. Marcamos el coeficiente correspondiente al enlace **d** de la Fig. 7.2.

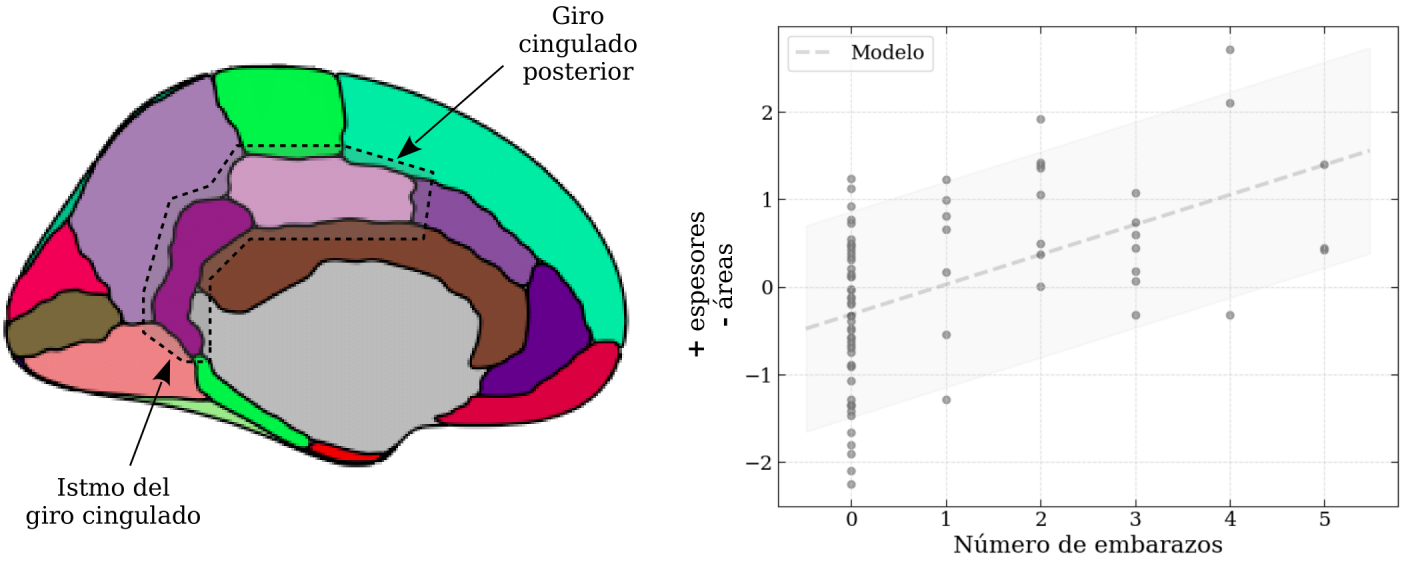

Figura 7.7: Estructuras pertenecientes a la comunidad de la variable latente del enlace **d** de la Fig. 7.2 (izquierda). Coordenada de cada sujeto sobre la dirección anatómica de la variable latente del enlace **d** (en desviaciones estándar) en función del número de embarazos (derecha). Mostramos la dependencia estimada por el modelo como una línea de trazo discontinuo con una franja gris que corresponde a la media ± una desviación estándar de la coordenada.

volumen intracraneal. Es decir que recuperamos la correlación entre el sexo y el tamaño cerebral que mostramos en la sección 2.2, donde encontramos que los hombres tienden a tener un cerebro más grande que las mujeres y mencionamos que, más que al sexo, tal relación se asocia a una diferencia de tamaño corporal.

La Fig. 7.4 muestra los coeficientes de la columna de A correspondientes a la altura para valores crecientes de γ_0. A medida que aumentamos γ_0, nos volvemos más exigentes para detectar una relación relevante, y en consecuencia, estimamos menos coeficientes distintos de cero. La influencia más estable por parte de la altura para valores altos de γ_0 es sobre la variable latente del enlace **c**. En la Fig. 7.5 mostramos la composición anatómica de tal variable latente (izquierda) y el valor que toma para cada sujeto en función de la altura (derecha), donde se aprecia que la distribución de puntos realmente disminuye con la altura. La combinación lineal de la variable latente revela que al movernos a lo largo de su correspondiente dirección anatómica los espesores del giro fusiforme y del giro parahipocampal tienden a aumentar y sus volúmenes tienden a disminuir.

La Fig. 7.6 muestra los coeficientes de la columna de A correspondientes al número de embarazos para valores crecientes de γ_0. La influencia más estable para valores altos de γ_0 es sobre la variable latente del enlace **d**. En la Fig. 7.7 mostramos la composición anatómica de tal variable latente (izquierda) y el valor que toma para cada sujeto en función del número de embarazos (derecha), donde se aprecia que la distribución de puntos realmente aumenta con el número de embarazos. La combinación lineal de la variable latente revela que al movernos a lo largo de su correspondiente dirección

anatómica los espesores de la región cingulada tienden a aumentar y sus áreas tienden a disminuir.

Capítulo 8

Conclusión

En este estudio realizamos un relevamiento de las propiedades neuroanatómicas de la población control utilizando técnicas de resonancia magnética nuclear, y de las características socioambientales de los voluntarios, a través de encuestas. Esto conllevó el desarrollo de métodos bayesianos para determinar la existencia de correlaciones entre variables anatómicas, y entre variables anatómicas y variables socioambientales. Reclutamos 77 voluntarios adultos, sin patologías evidentes, residentes en San Carlos de Bariloche y zona de influencia. Analizamos nuestras imágenes T1 en conjunto con las recolectados por la unidad ejecutora de Estudios en Neurociencias y Sistemas Complejos (ENyS), obteniendo una muestra de $n = 193$ individuos.

Mostramos que los resultados de pruebas estadísticas dependen fuertemente del espacio elegido para representar los datos. Si uno no realiza una reducción de dimensión, puede suceder que el gran número de dimensiones muestreadas enmascare un pequeño número de dimensiones que revelan la diferencia entre dos poblaciones con distintas características socioambientales. Sin embargo, la reducción de la dimensión no puede hacerse únicamente con el objetivo de encontrar una diferencia entre poblaciones, porque este procedimiento da lugar a distinciones entre poblaciones sólo como consecuencia de muestreo finito.

En el trabajo de licenciatura redujimos la dimensión mediante un análisis de componentes principales aprovechando las correlaciones poblacionales que existen entre las medidas antómicas. Identificamos a las dos primeras componentes principales como direcciones de tamaño global y de espesor de medidas corticales respectivamente. Al mismo tiempo, encontramos que la primera componente está relacionada con la altura (que a su vez está correlacionada con el sexo) y la segunda con la edad.

Durante la maestría diseñamos un procedimiento de reducción de dimensión que nos permitió dividir las estructuras anatómicas en comunidades, con correlaciones fuertes dentro de cada comunidad, y débiles entre comunidades. Observamos que el número de comunidades depende del tamaño de la población muestreada, detectando la existencia

de un tamaño crítico, por debajo del cual es imposible segmentar el grafo de correlaciones directas. Transiciones de este tipo fueron observadas anteriormente en trabajos que describen la física estadística de la inferencia [45].

Las comunidades resultantes son: (a) una comunidad con regiones ocupadas por líqudo encefalorraquídeo y materia blanca con conexiones de largo alcance, como el cuerpo calloso y el quiasma óptico, (b) una comunidad con regiones subcorticales y corticales filogenéticamente antiguas, y (c) 21 comunidades corticales, en su mayoría conteniendo estructuras homólogas de ambos hemisferios cerebrales y funcionalmente conectadas. Dentro de cada comunidad cortical, el volumen de una dada región típicamente se correlacionaba positivamente con los correspondientes áreas y espesores, mientras que el área y el espesor se correlacionaban negativamente. Las estructuras a uno y otro lado del cerebro típicamente se conectaban a través de los espesores y volúmenes de materia blanca. El algoritmo de reducción de la dimensión determinó 41 variables latentes, mayormente conteniendo áreas y volúmenes. Las variables latentes se eligieron comparando la variabilidad inter-sujeto con la intra-sujeto. Cada variable latente pertenece a una única comunidad, lo cual permite localizarla en una región específica del cerebro - a diferencia de las componentes principales obtenidas por PCA.

Posteriormente determinamos el grafo de conexiones entre pares de variables, tanto del tipo latente - latente (matriz Ω) como socioambiental - latente (matriz A). El algoritmo implementado es conservador. Por lo tanto, no podemos descartar que existan correlaciones adicionales que no estamos detectando. Lo que podemos asegurar es que aquello que no detectamos es menos relevante que lo que sí detectamos. El procedimiento nos permite contar con una perilla (γ_0), con la que podemos regular dónde hacer el corte de significancia.

Encontramos que los efectos socioambientales más robustos involucran al sexo, la altura y el número de embarazos. En el caso del sexo, recuperamos la correlación con el tamaño global del cerebro que reportamos en el trabajo de licenciatura. Estimamos que la altura tiene un efecto sobre la geometría del giro fusiforme y del giro parahipocampal. Al mismo tiempo, el número de embarazos correlaciona fuertemente con medidas anatómicas correspondientes a la región cingulada del cerebro.

El modelo de inferencia desarrollado en esta tesis tiende a explicar la variabilidad de los datos describiendo únicamente las interacciones más relevantes entre variables. En consecuencia, podría no capturar los efectos de variables socioambientales que tienen una influencia pequeña sobre un gran número de medidas anatómicas. Esta puede ser la razón por lo que no recuperamos el efecto de la edad sobre los espesores encontrado en la licenciatura. Trabajo futuro podría consistir en hacer el análisis únicamente con las medidas de espesor. Al mismo tiempo, se podría analizar por separado los datos de cada resonador, con el objetivo de determinar si tiene alguna influencia, y en caso afirmativo, identificar si el factor relevante es la etnia de distintas regiones geográficas,

o las características del equipo. Por último, a partir de los datos recolectados se podría confeccionar un atlas que caracterice la neuroanatomía de nuestra región geográfica, así como su dependencia con los factores socioambientales, que sirva de referencia para estudios futuros.

Apéndice A

Encuesta socioambiental

ENCUESTA

DATOS PERSONALES:

Nombre							
Apellido							
DNI							
Teléfono							
Email							
Fecha de nacimiento	Día		Mes		Año		
Edad							
Sexo	Femenino		Masculino		No contesto		
Estatura							
Peso							
Mano dominante	Derecha		Izquierda		Ambas		

ORIGEN GEOGRÁFICO:

Lugar de nacimiento	
¿Dónde vivió los primeros 6 años de su vida?	
Lugar de residencia actual	
¿Hace cuánto vive en su lugar de residencia actual?	

LUGAR DE NACIMIENTO DE LOS PADRES Y ABUELOS:

(Si no conoce la ciudad de nacimiento indique la provincia y/o el país)

Lugar de nacimiento de...

madre biológica	
padre biológico	
abuela materna	
abuelo materno	
abuela paterna	
abuelo paterno	

HISTORIA DE VIDA:

(Las preguntas sobre sus hermanos se refieren a aquellos con los que comparta misma madre. Incluya tanto los que están vivos como los que han fallecido.)

Peso al nacimiento	

Edad (en meses) al nacimiento	9 meses		7-8 meses		No sé	
¿Fue amamantado?	Sí			No		
¿Cuántos hermanos tiene?						
¿Tiene hermanos gemelos/mellizos?	Sí			No		
¿En qué orden nació usted con respecto a sus hermanos?						
Número de embarazos						
Número de hijos nacidos vivos						
Edad a la que tuvo su primer hijo						

EDUCACIÓN:

Nivel educativo alcanzado	Primario incompleto	
	Primario completo	
	Secundario incompleto	
	Secundario completo	
	Terciario/Universitario incompleto	
	Terciario/Universitario completo	

Si adquirió o está adquiriendo habilidades específicas (deportivas, artísticas, terciarias/universitarias o especializaciones) indique nombre y tiempo de formación:

Carrera/Disciplina	Tiempo de formación

Nivel educativo de la madre	Primario incompleto	
	Primario completo	
	Secundario incompleto	
	Secundario completo	
	Terciario/Universitario incompleto	
	Terciario/Universitario completo	

Nivel educativo del padre	Primario incompleto	
	Primario completo	
	Secundario incompleto	
	Secundario completo	

	Terciario/Universitario incompleto	
	Terciario/Universitario completo	

¿Habla, lee o comprende un idioma además del español?	Sí		No	
Si su respuesta anterior fue sí, indique cuál/es				
Idioma original que habla su madre				

LABORAL:

Ocupación	
Ocupación de la madre	
Ocupación del padre	

SALUD:

	Sí	No
¿Es hipertenso?		
¿Es diabético?		
¿Es asmático?		
¿Tiene antecedentes alérgicos de importancia?		
¿Toma antiinflamatorios diarios?		
¿Tiene antecedentes familiares de enfermedad de Alzheimer?		
¿Tiene problemas de memoria con frecuencia?		
¿Tiene antecedentes de trauma de cráneo o columna que hayan motivado internación?		
¿Toma corticoides con frecuencia?		
¿Fuma? (aclare cantidad de cigarillos diarios)		
¿Toma estatinas? (atorvastatina, simvastatina, rosuvastatina, cerivastatina)		
¿Tiene antecedentes personales de enfermedad cardiovascular?		
¿Tiene epilepsia?		

Responda si posee alguno de los siguientes items:

	Sí	No
Cápsula para endoscopía de intestino delgado		

Desfibrilador cardiaco implantado (anterior o actual)		
Dispositivo de asistencia para el ventrículo izquierdo (LVAD) (bomba cardiaca)		
Expansores del tejido mamario		
Marcapasos o cables para marcapasos (anteriores o actual)		
Neuroestimulador implantado		
Embarazo		
Grapas para aneurismas		
Colonoscopia reciente o procedimiento del sistema digestivo que incluye grapas quirúrgicas		
Grapas quirúrgicas/grapas vasculares/injertos		
Bomba para medicamentos		
Unidad de neuroestimulación eléctrica transcutánea		
Cuerpo metálico extraño (heridas de bala, cerclaje de retina)		
Lesión ocular que incluye metal		
Cirugía previa (craneal, cardíaca, ocular, etc.). Detalle:		
Tratamiento para la gota		
Válvulas cardiacas artificiales/endoprótesis coronarias		
Antecedentes de caídas durante los últimos 30 días		
Asma o enfermedad coronaria crónica, insuficiencia cardiaca congestiva (CHF)		
Filtro en forma de paraguas en la vena cava		
Reemplazos de articulaciones/implantes. Detalle:		
Dispositivos ortopédicos o protésicos		
Claustrofobia		
Insuficiencia renal/hepática/cardíaca		
Extensiones de cabello/postizo/peluca		
Aparatos de ortodoncia o resortes orales		
Prótesis dentales removibles		
Brillo/maquillaje de ojos permanentes		
Tatuajes o perforaciones en el cuerpo		
Audífonos removibles		
Parches para la piel con medicamento		
Ferropenia tratada con Feraheme		

Bibliografía

[1] Davison Ankney, C. Sex differences in relative brain size: The mismeasure of woman, too? *Intelligence*, **16** (3), 329 – 336, 1992. URL http://www.sciencedirect.com/science/article/pii/016028969290013H, special Issue: Biology and Intelligence. 1

[2] Ling, S., Umbach, R., Raine, A. Biological explanations of criminal behavior. *Psychology, Crime & Law*, **25** (6), 626–640, 2019. 1

[3] Reuter, M., Rosas, H. D., Fischl, B. Highly accurate inverse consistent registration: A robust approach. *NeuroImage*, **53** (4), 1181–1196, 2010. URL http://dx.doi.org/10.1016/j.neuroimage.2010.07.020. 4

[4] Segonne, F., Dale, A. M., Busa, E., Glessner, M., Salat, D., Hahn, H. K., *et al.* A hybrid approach to the skull stripping problem in mri. *NeuroImage*, **22** (3), 1060 – 1075, 2004. URL http://www.sciencedirect.com/science/article/B6WNP-4CF5CNY-1/2/33cc73136f06f019b2c11023e7a95341. 4

[5] Sled, J., Zijdenbos, A., Evans, A. A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE Trans Med Imaging*, **17**, 87–97, 1998. 4

[6] Fischl, B., Sereno, M. I., Dale, A. Cortical surface-based analysis: Ii: Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, **9** (2), 195 – 207, 1999. 4

[7] Dale, A., Fischl, B., Sereno, M. I. Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage*, **9** (2), 179 – 194, 1999.

[8] Fischl, B., Sereno, M. I., Tootell, R. B., Dale, A. M. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human Brain Mapping*, **8** (4), 272–284, 1999. URL http://dx.doi.org/10.1002/(SICI)1097-0193(1999)8:4<272::AID-HBM10>3.0.CO;2-4. 4

[9] Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., *et al.* An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *NeuroImage*, **31** (3), 968 – 980, 2006. URL http://www.sciencedirect.com/science/article/B6WNP-4JFHF4P-1/2/0ec667d4c17eafb0a7c52fa3fd5aef1c. 4

[10] Jaynes, E. T. Prior probabilities. *IEEE Transactions on systems science and cybernetics*, **4** (3), 227–241, 1968. 18

[11] Jaynes, E. T. Information theory and statistical mechanics. *Physical review*, **106** (4), 620, 1957. 18

[12] Moon, T. K. The expectation-maximization algorithm. *IEEE Signal processing magazine*, **13** (6), 47–60, 1996. 22

[13] Williams, D. R., Rast, P. Back to the basics: Rethinking partial correlation network methodology. *British Journal of Mathematical and Statistical Psychology*, **73** (2), 187–212, 2020. 22

[14] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, **2008** (10), P10008, 2008. 29

[15] Dapretto, M., Bookheimer, S. Form and content dissociating syntax and semantics in sentence comprehension. *Neuron*, **24**, 427–432, 1999. 32

[16] Dronkers, N., Wilkins, D., Valin, R. V., Redfern, B. B., Jaeger, J. Lesion analysis of the brain areas involved in language comprehension. *Cognition*, **92**, 145–177, 2004. 32

[17] Sprengelmeyer, R., Rausch, M., Eysel, U. T., Przuntek, H. Neural structures associated with recognition of facial expressions of basic emotions. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **265** (1409), 1927–1931, 1998. 32

[18] Hafting, T., Fyhn, M., Molden, S., Moser, M., Moser, E. Microstructure of a spatial map in the entorhinal cortex. *Nature*, **436**, 801–806, 2005. 32

[19] Kropff, E., Carmichael, J., Moser, M., Moser, E. Speed cells in the medial entorhinal cortex. *Nature*, **523**, 419–424, 2015.

[20] Tsao, A., Sugar, J., Lu, L., Wang, C., Knierim, J. J., Moser, M.-B., *et al.* Integrating time from experience in the lateral entorhinal cortex. *Nature*, **561** (7721), 57–62, 2018. 32

[21] Vanni, S., Tanskanen, T., Seppä, M., Uutela, K., Hari, R. Coinciding early activation of the human primary visual cortex and anteromedial cuneus. *Proceedings of the National Academy of Sciences*, **98** (5), 2776–2780, 2001. 32

[22] Haldane, M., Cunningham, G., Androutsos, C., Frangou, S. Structural brain correlates of response inhibition in bipolar disorder i. *Journal of Psychopharmacology*, **22** (2), 138–143, 2008. 32

[23] Caplan, D. Why is broca's area involved in syntax? *Cortex*, **42** (4), 469–471, 2006. 33

[24] Grewe, T., Bornkessel, I., Zysset, S., Wiese, R., von Cramon, D. Y., Schlesewsky, M. The emergence of the unmarked: A new perspective on the language-specific function of broca's area. *Human brain mapping*, **26** (3), 178–190, 2005.

[25] Rodd, J. M., Davis, M. H., Johnsrude, I. S. The neural mechanisms of speech comprehension: fmri studies of semantic ambiguity. *Cerebral Cortex*, **15** (8), 1261–1269, 2005. 33

[26] Fadiga, L., Craighero, L., Destro, M. F., Finos, L., Cotillon-Williams, N., Smith, A., *et al.* Language in shadow. *Social Neuroscience*, **1**, 77 – 89, 2006. 33

[27] Fadiga, L., Craighero, L. Hand actions and speech representation in broca's area. *Cortex*, **42**, 486–490, 2006. 33

[28] Ghez, C. Fahn s. the cerebellum. *Principles of neural science*, 1985. 33

[29] Rapp, B. Handbook of cognitive neuropsychology: What deficits reveal about the human mind. Psychology Press, 2015.

[30] Doya, K. Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology*, **10**, 732–739, 2000.

[31] Hernaez-Goni, P., Tirapu-Ustarroz, J., Iglesias-Fernandez, L., Luna-Lario, P. The role of the cerebellum in the regulation of affection, emotion and behaviour. *Revista de neurologia*, **51** (10), 597, 2010.

[32] Turner, B. M., Paradiso, S., Marvel, C. L., Pierson, R., Ponto, L. L. B., Hichwa, R. D., *et al.* The cerebellum and emotional experience. *Neuropsychologia*, **45** (6), 1331–1341, 2007. 33

[33] Hikosaka, O., Takikawa, Y., Kawagoe, R. Role of the basal ganglia in the control of purposive saccadic eye movements. *Physiological reviews*, **80 3**, 953–78, 2000. 34

[34] Ikemoto, S., Yang, C., Tan, A. Basal ganglia circuit loops, dopamine and motivation: a review and enquiry. *Behavioural brain research*, **290**, 17–31, 2015.

[35] Redgrave, P., Prescott, T., Gurney, K. The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience*, **89**, 1009–1023, 1999.

[36] Schroll, H., Hamker, F. H. Computational models of basal-ganglia pathway functions: focus on functional neuroanatomy. *Frontiers in systems neuroscience*, **7**, 122, 2013. 34

[37] et de Physiologie, T. d. Limbic system. 34

[38] Olds, J., Milner, P. Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. *Journal of comparative and physiological psychology*, **47** (6), 419, 1954.

[39] Adams, R. D. Principles of neurology. McGraw-Hill, Health Professions Division, 1997. 34

[40] Aggleton, J. P., Dumont, J. R., Warburton, E. C. Unraveling the contributions of the diencephalon to recognition memory: a review. *Learning & memory*, **18** (6), 384–400, 2011. 34

[41] Bishop, C. M. Bayesian pca. En: Advances in neural information processing systems, págs. 382–388. 1999. 35, 37, 39

[42] Tipping, M. E., Bishop, C. M. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **61** (3), 611–622, 1999. 36

[43] Meng, X.-L., Rubin, D. B. Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, **80** (2), 267–278, 1993. 46

[44] Deshpande, S. K., Ročková, V., George, E. I. Simultaneous variable and covariance selection with the multivariate spike-and-slab lasso. *Journal of Computational and Graphical Statistics*, **28** (4), 921–931, 2019. 46

[45] Zdeborová, L., Krzakala, F. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, **65** (5), 453–552, 2016. 54

Agradecimientos

Quiero agradecer a los colaboradores de este estudio: Sergio, Mariana V, Mariana B, Paula, Silvia y Juan Pablo. A partir nuestras discusiones se generaron muy buenas ideas que terminaron siendo implementadas en el análisis y que a mí sólo nunca se me hubieran ocurrido. En particular, también agradezco a Sergio y a Juan Pablo por tomarse el trabajo de mirar tantas resonancias.

Muchas gracias a las personas que manejaron el resonador del Centro de Radioterapia y Medicina Nuclear: Humberto, Virginia y Florencia. Sin ustedes no hubiéramos podido escanear ningún voluntario.

Agradezco enormemente a todos los voluntarios que participaron en este estudio por su tiempo y su buena onda, sobre todo a aquellos valientes que aceptaron escanearse múltiples veces.

Muchas gracias también al grupo de compneuro, en particular a Damián, Nico y Seba, tanto por las ideas que aportaron en las reuniones de grupo como por hacerme pasar lindos momentos, por ejemplo destruyéndome en el pictionary.

Muchas gracias a mis compañeros y amigos. Siempre voy a recordar mi tiempo en el IB como una linda experiencia, y ustedes son un motivo fundamental para que lo haya disfrutado tanto.

Por último, muchas gracias a mi directora, Inés. Gracias por acompañarme este año y medio de maestría, tanto en aspectos de trabajo como personales. Gracias por todo el tiempo que me dedicaste. Gracias por tantas discusiones en las que no solo aportabas tus lindas ideas, sino que además volvías lindas las mías. Gracias por acompañarme todos los sábados de resonancias, y traer mate y tus MUY ricas tortas. Gracias por tus excelentes clases, ranro presenciales como virtuales. Gracias por organizar muy lindas actividades grupales (estaría aún más agradecido si mi muñeco de nieve hubieses ganado la caja de chocolates). Gracias por formarme en nueve dimensiones. Muchas gracias por tomarme como estudiante, este año y medio de maestría fue súper lindo.